Learning the language of smell: Foundation models for protein–odor interactions

BRIAN DEPASQUALE

ASSISTANT PROFESSOR, DEPT. OF BIOMEDICAL ENGINEERING

AFFILIATE FACULTY, HARIRI INSTITUTE FOR COMPUTING

BOSTON UNIVERSITY

NEW(-ISH) LAB AT BU

New(-ish) lab at BU!

Grant McConachie
Darcy Zi
Ryan Senne
Brittany Ahn
Omar El Sayed
Zach Loschinskey
Tushar Arora
Halley Dante

Collaborations

Ben Scott (BU)
Meg Younger (BU)
Mike Economo (BU)
Mark Howe (BU)

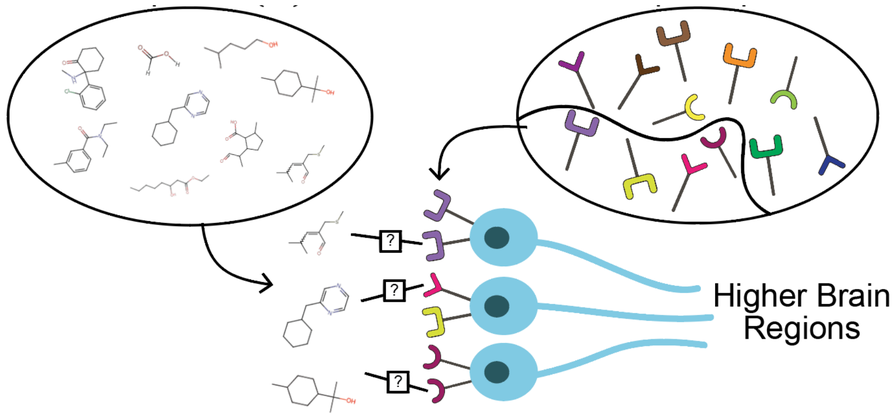
Past Collaborators

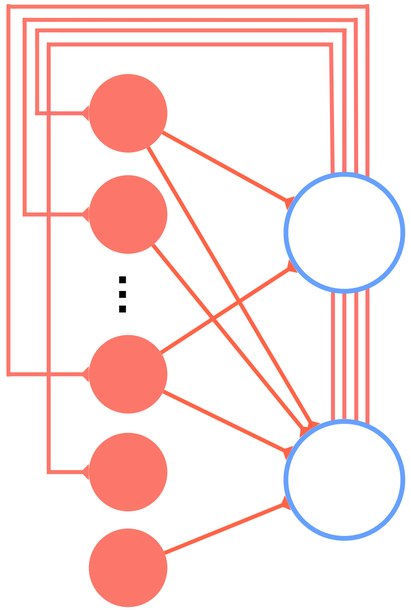
Larry Abbott (Columbia)
Mark Churchland (Columbia)
David Sussillo (Stanford)
Jonathan Pillow (Princeton)
Carlos Brody (Princeton)

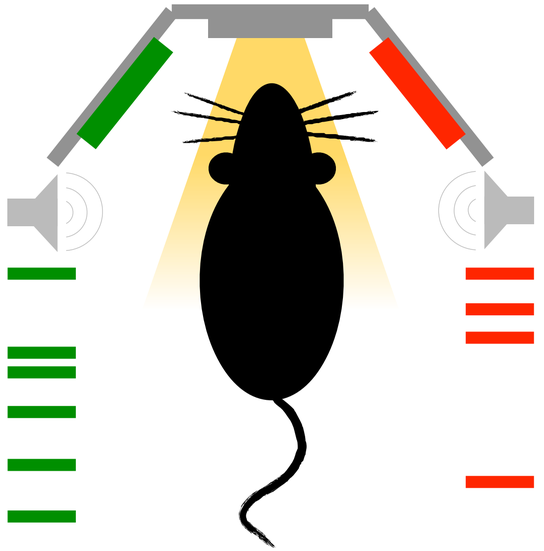
THE DEPAQ LAB @ BU

BRAIN 🧠 ⟷ BEHAVIOR 🕺

ML for biochemistry ⟵ Neural circuits ⟶ Models of decision-making





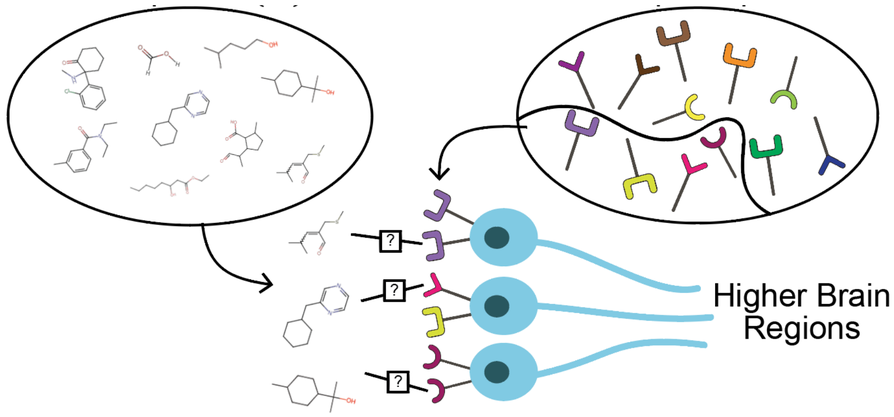


Abbott, DePasquale, et al. 2016
DePasquale, Churchland, Abbott, 2016
DePasquale et al. 2018
Pinto, Rajan, DePasquale et al. 2019
DePasquale, Sussillo, Abbott, Churchland, 2023

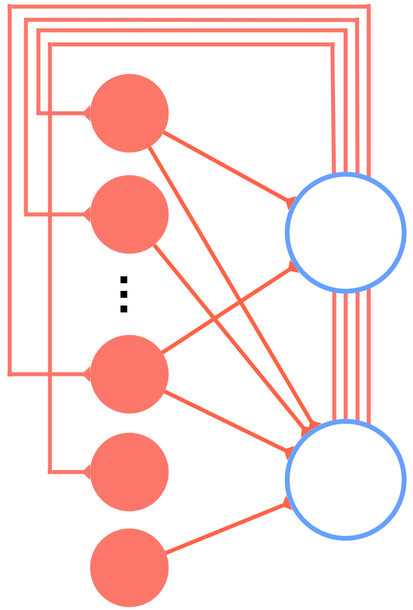
DePasquale, Brody, Pillow. 2024 eLife

THE DEPAQ LAB @ BU

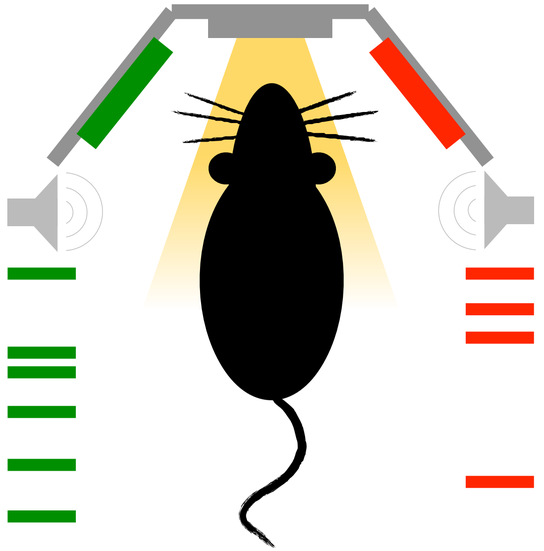
Foundation models for olfaction



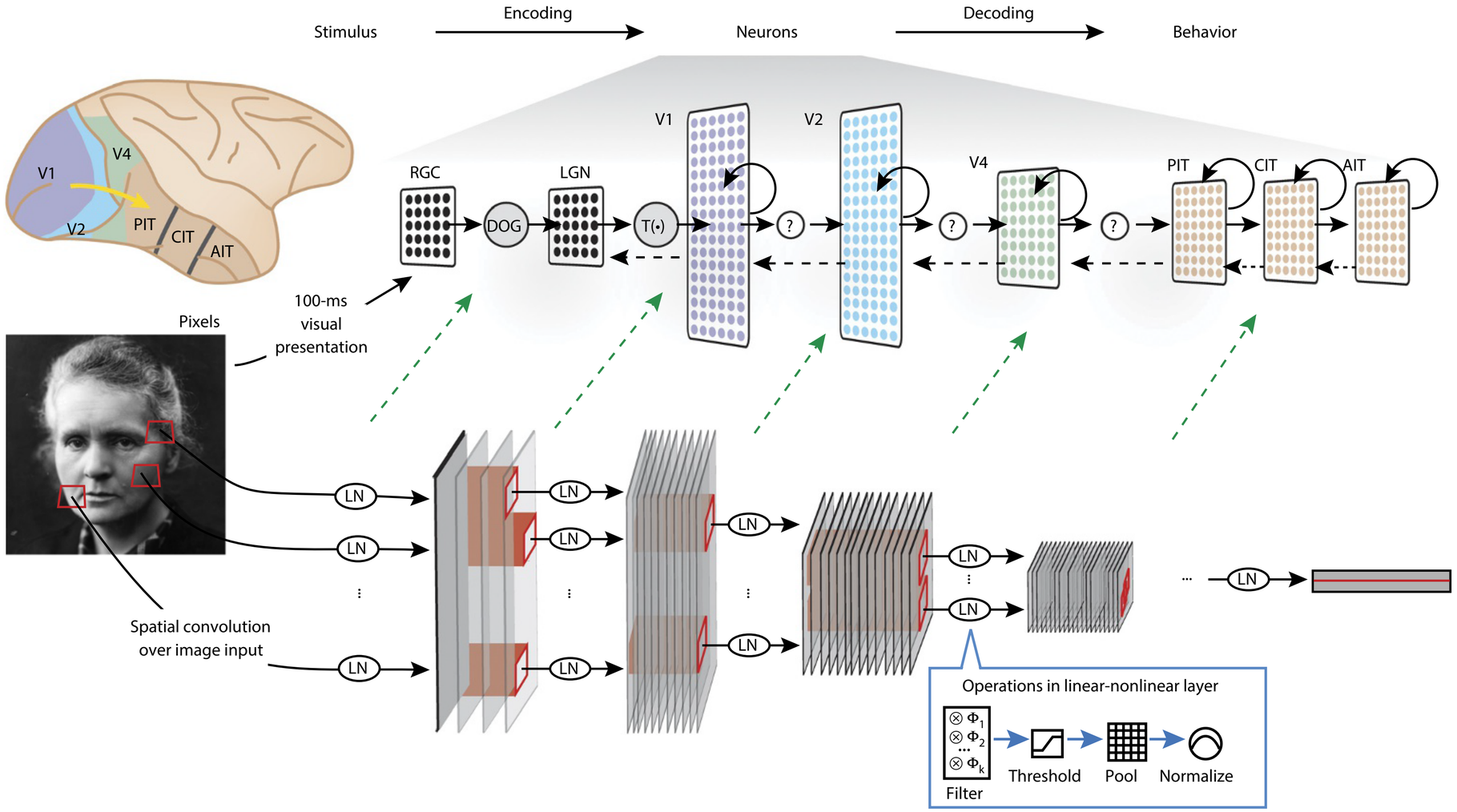
Spike variability from multi-task spiking networks



Hierarchical models of decision-making

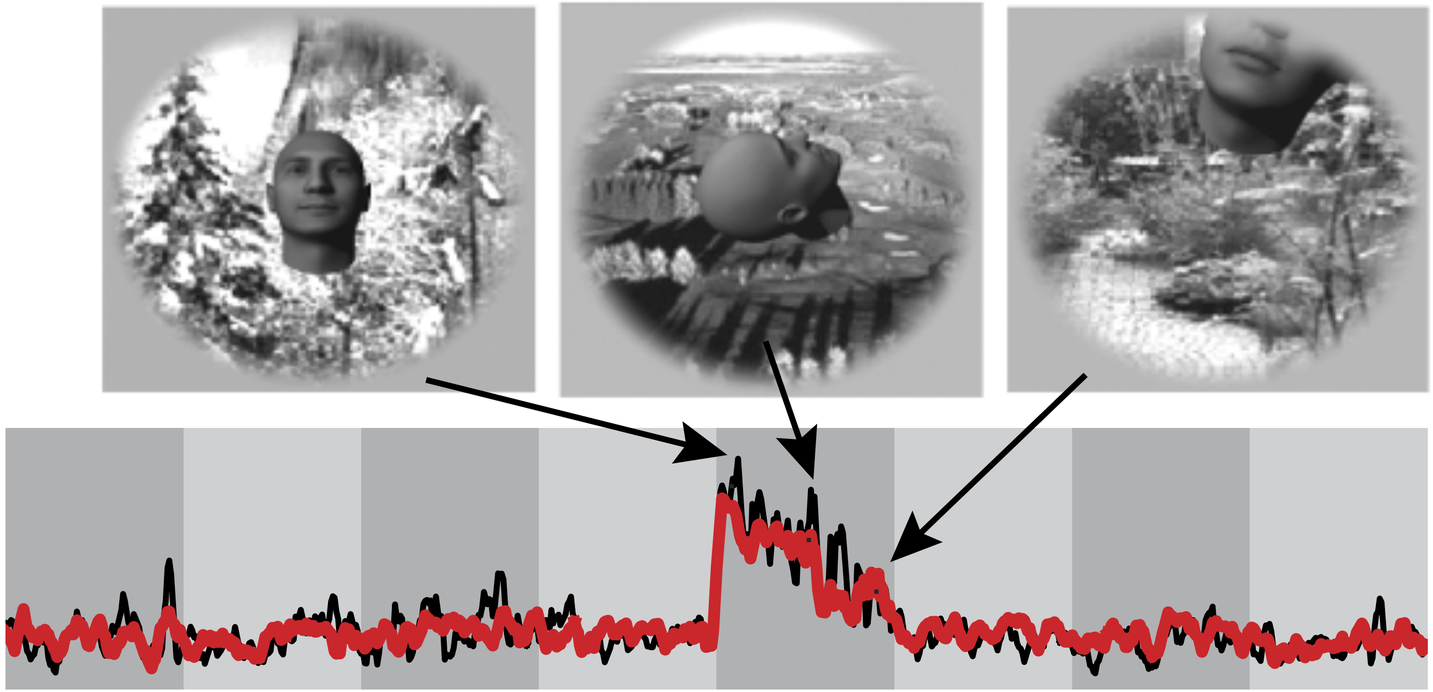


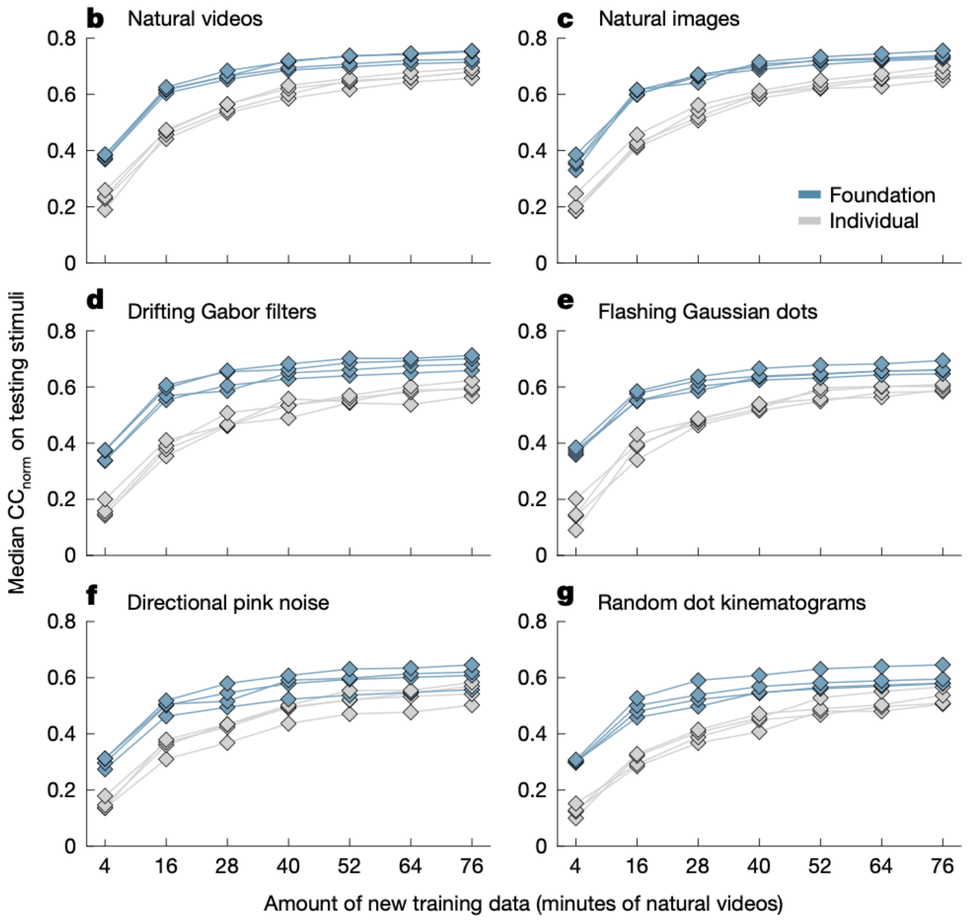
Vision has been a huge success story



Yamins, 2016

Vision has been a huge success story



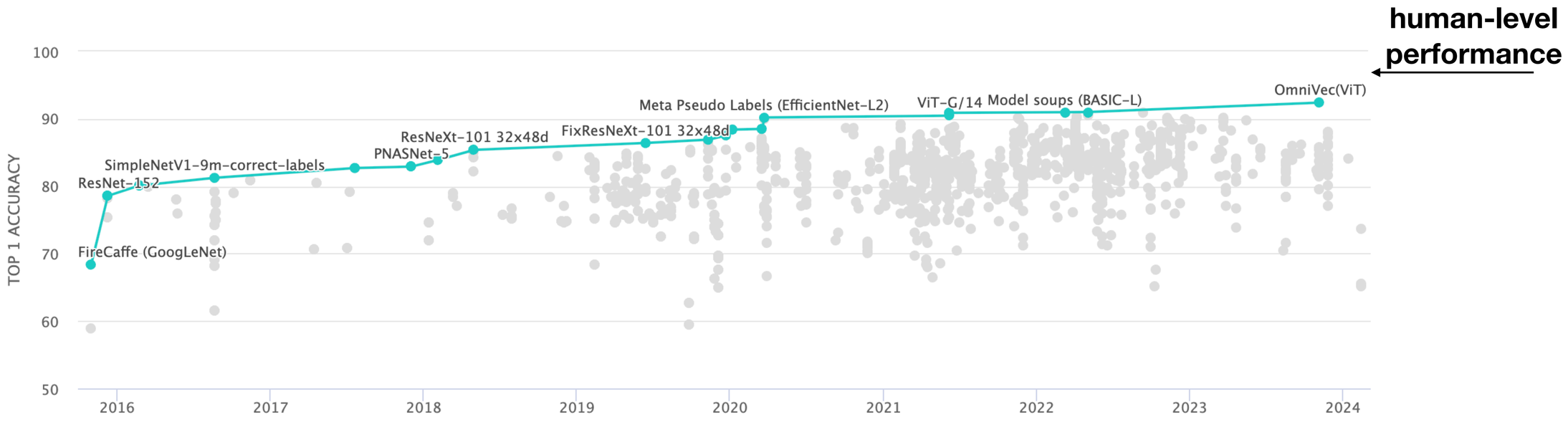


Foundation model of neural activity predicts response to new stimulus types

Yamins, 2016

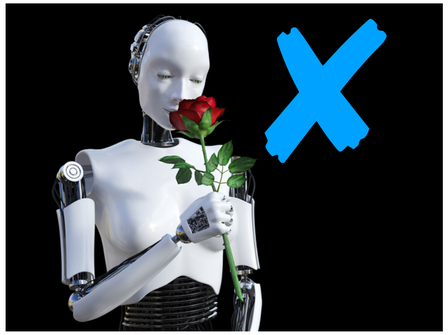
Wang, 2025

Vision has been a huge success story



…less so in olfaction

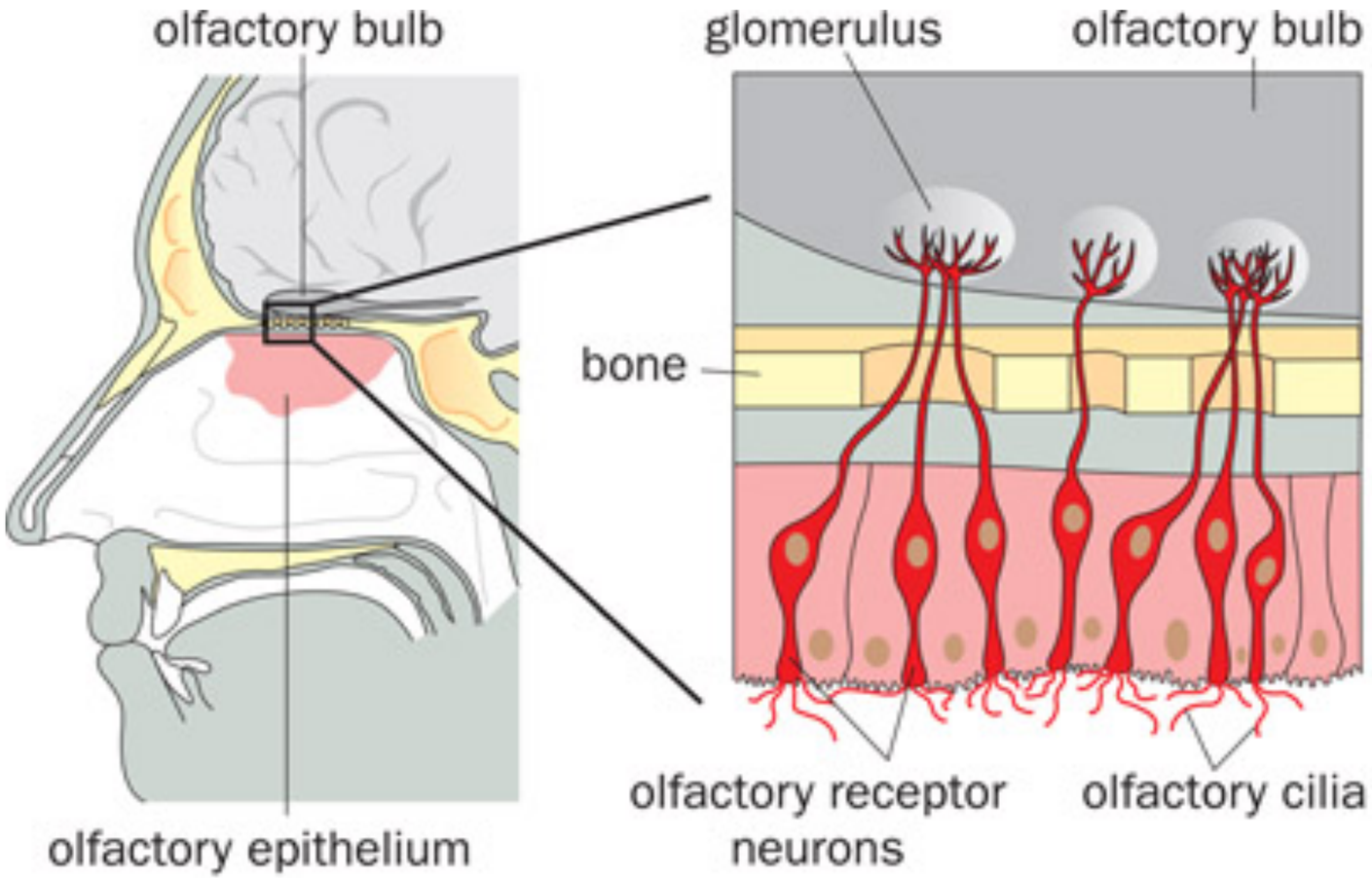
fruity or grassy?



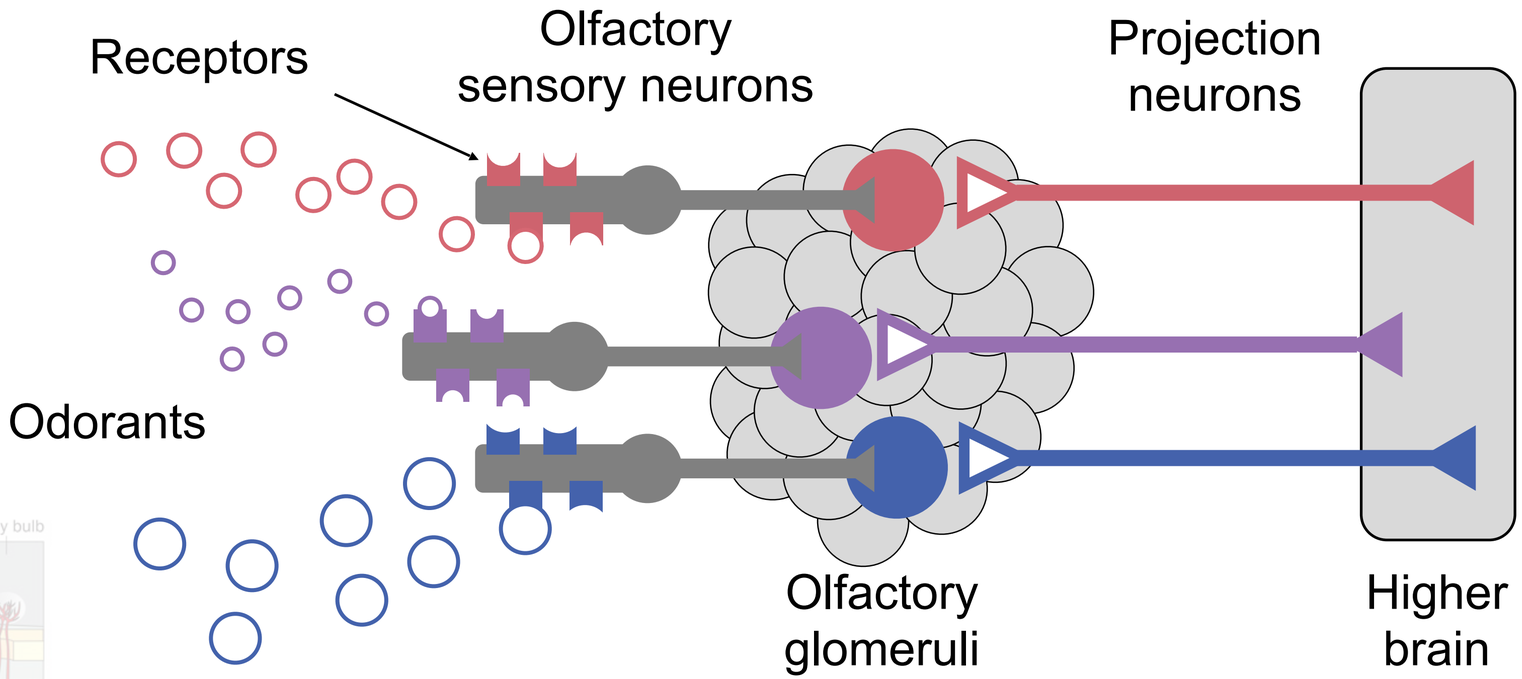
Top dog (best model)

Great ML models in biochemistry that are ripe for olfaction and/or neuroscience

How do *I* smell?

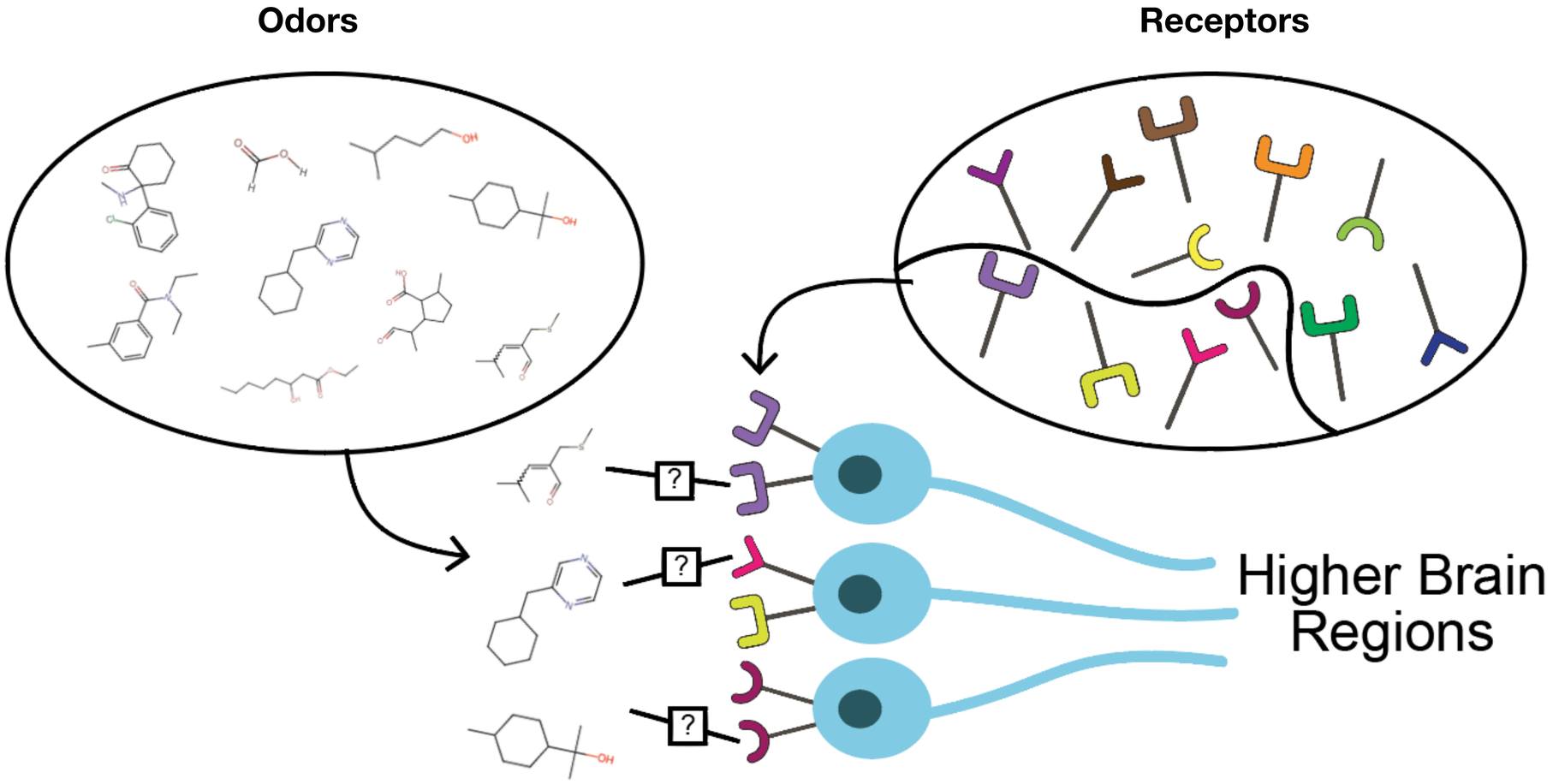


How do *I* smell?



How do *we* smell?

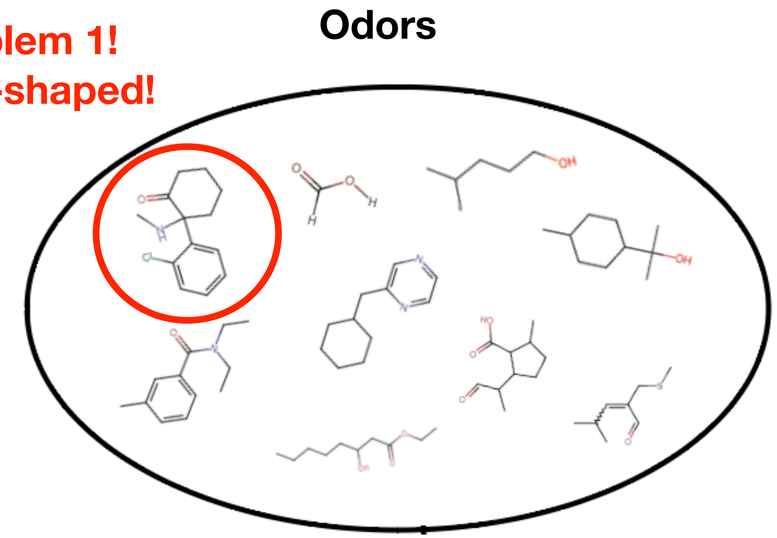
What's the problem?

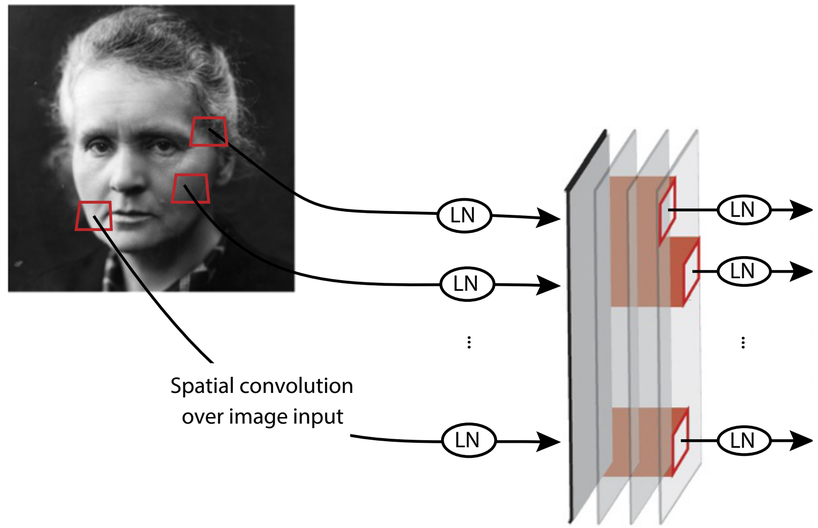


How do *we* smell?

What's the problem?

Problem 1!
Oddly-shaped!



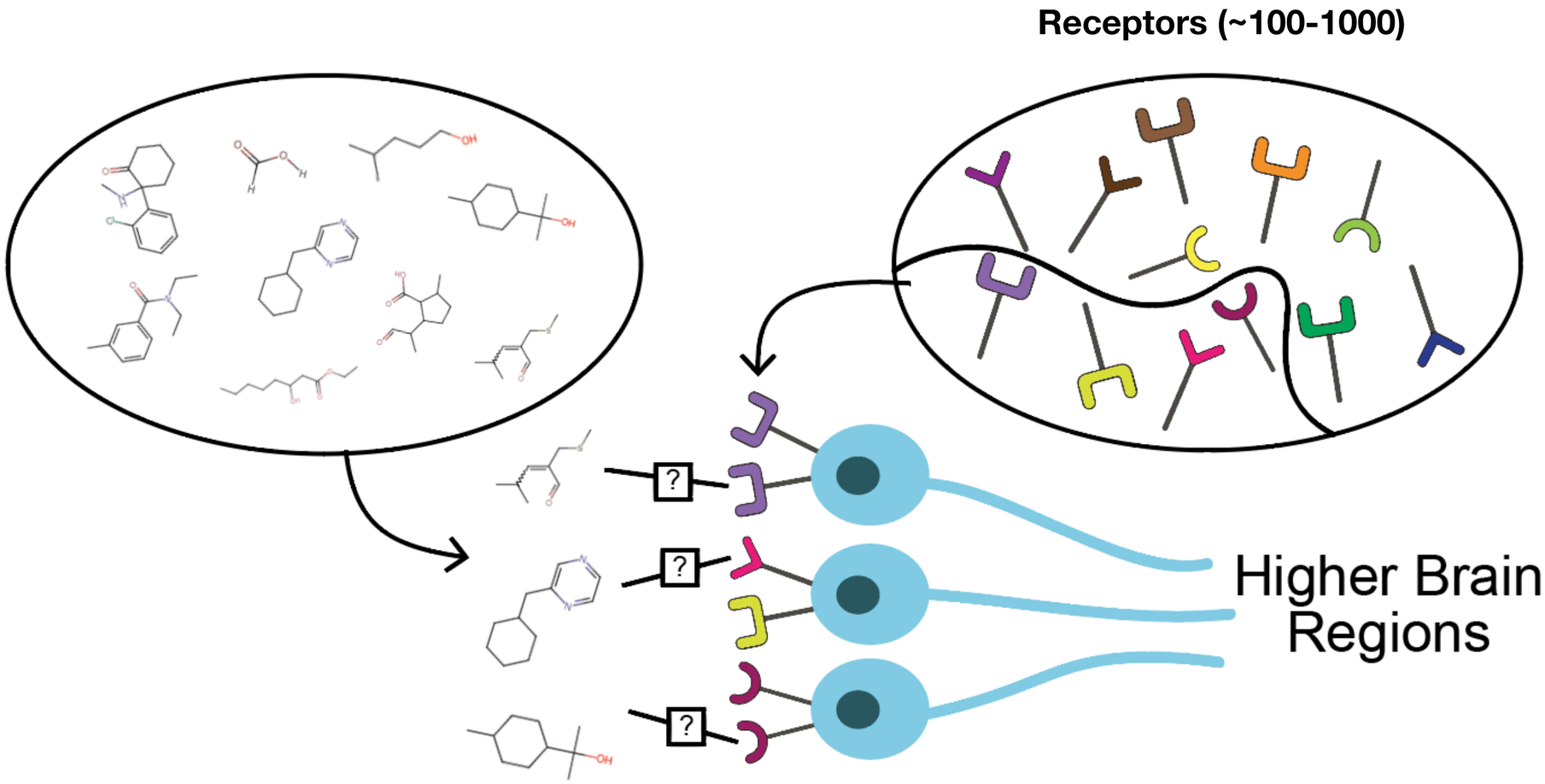


How do *we* smell?

Problem 2

What's the problem?

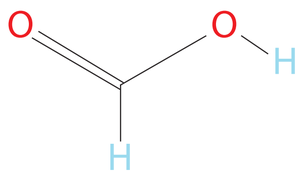
Dataset ~100-1000 😩



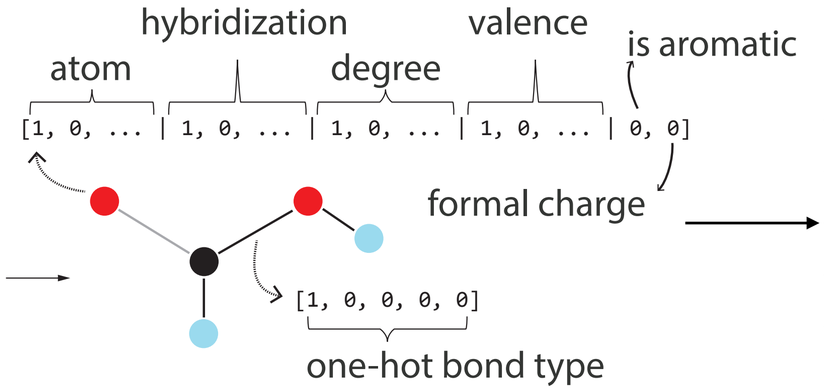
Mayhew et. al. (2022)

P1: Dealing with odd shapes

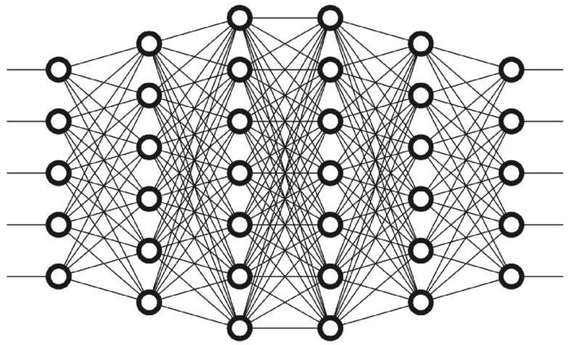
Start with chemicals



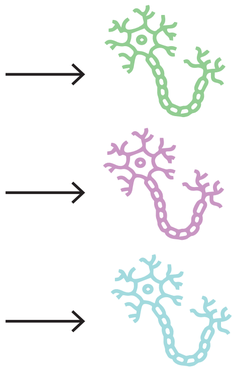
Turn into "featurized" graphs



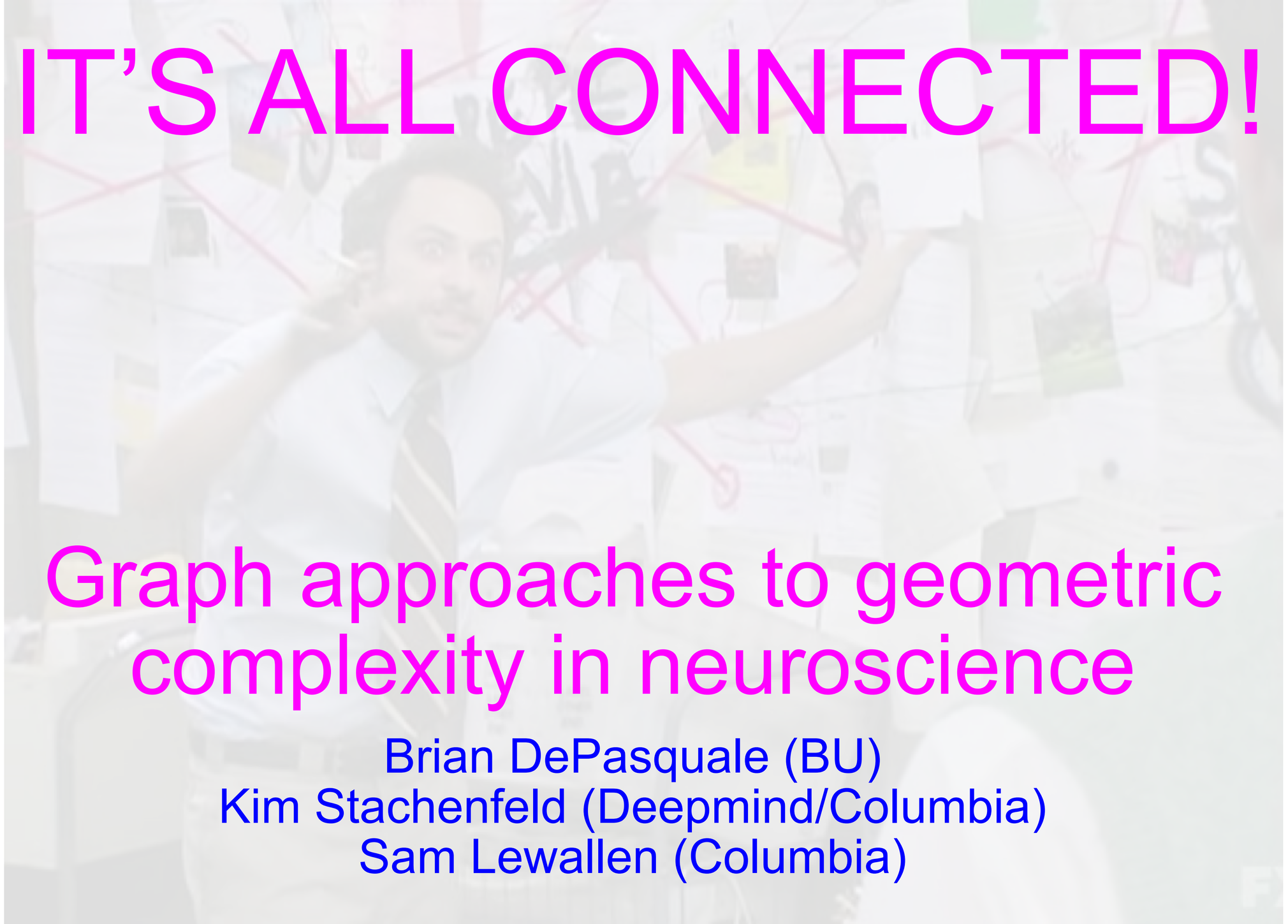
Pass nodes and edges through a neural network



Predict neural responses



w/ Grant McConachie



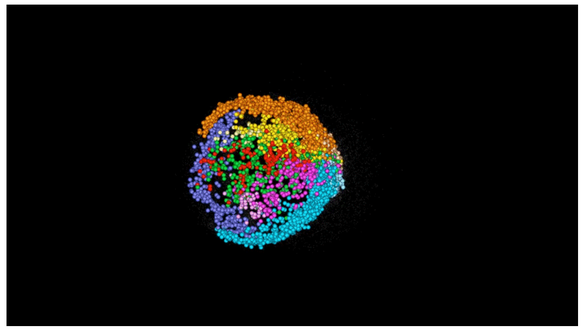
IT'S ALL CONNECTED!

Graph approaches to geometric complexity in neuroscience

Brian DePasquale (BU)
Kim Stachenfeld (Deepmind/Columbia)
Sam Lewallen (Columbia)

Geometrically complex data are *everywhere* in biology

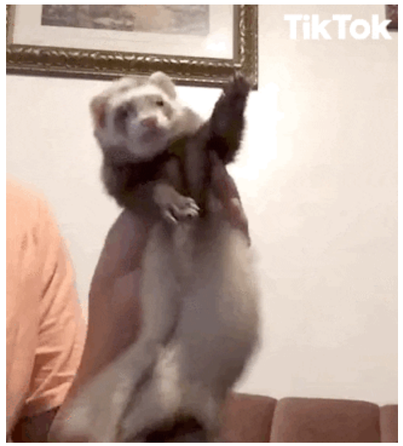
Development



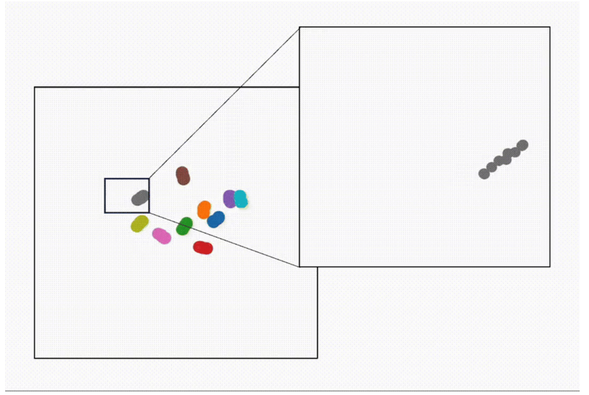
McDole et al 2018

Neural Dynamics

Biomechanics

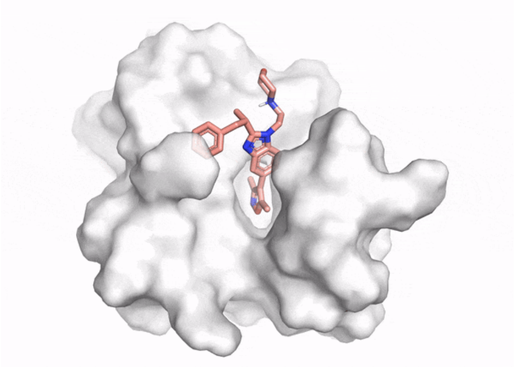


Multiagent behavior



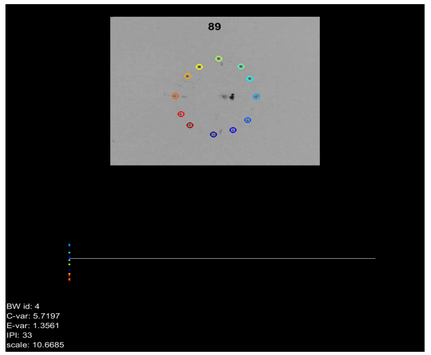
Courtesy of Matt Lovett-Barron (UCSD) and Grant McConachie (BU)

Biochemistry

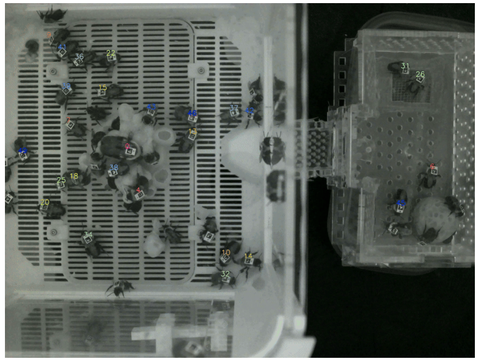


Connectomics

Behavior keypoints



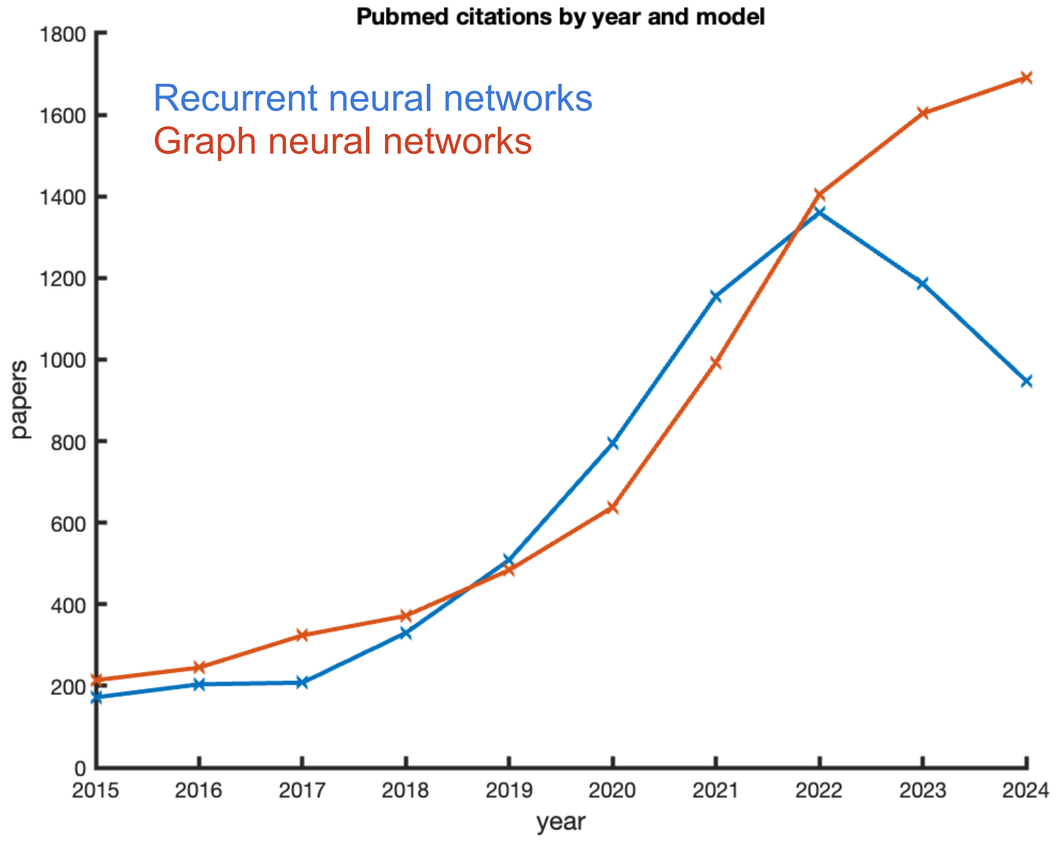
Courtesy of Brady Weissbourd (MIT)

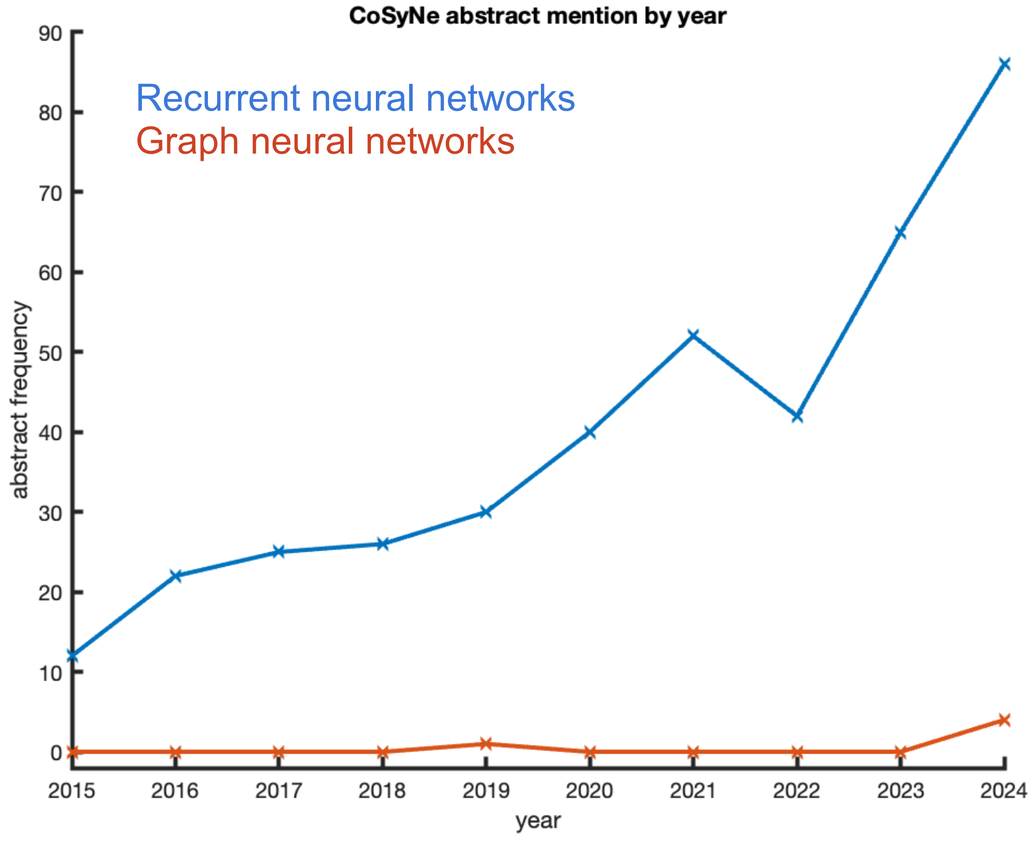


Courtesy of James Crall (Wisc-Mad)

Organized with Kim Stachenfeld and Sam Lewallen

Something is rotten in the state of CoSyNe

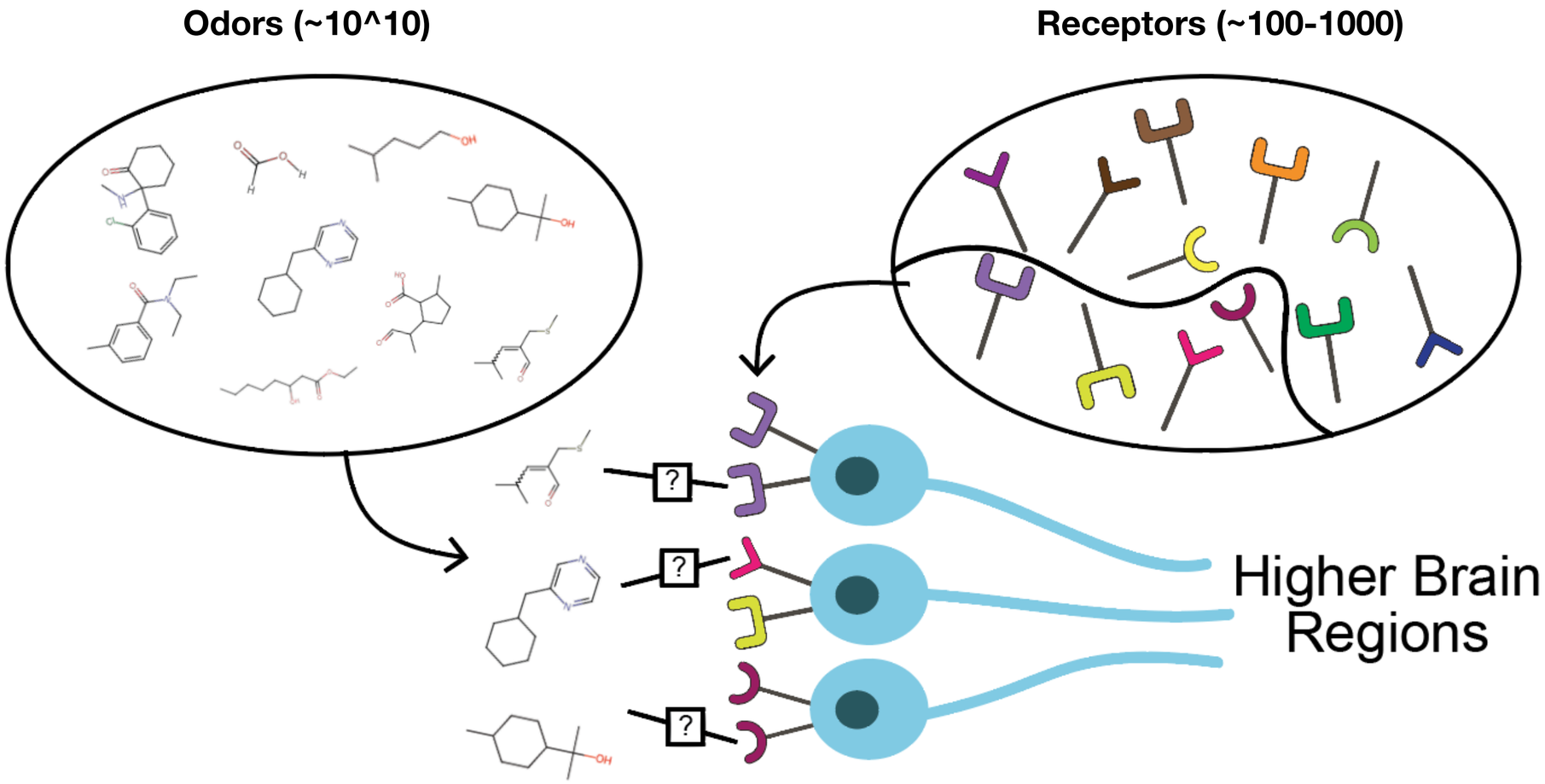




How do we smell?

Problem 2

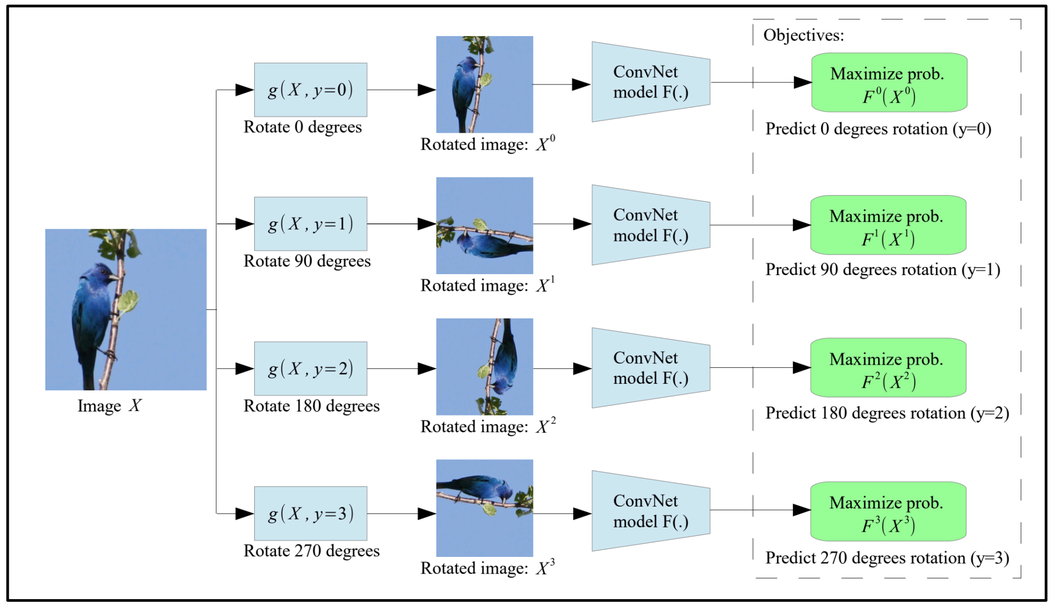
Dataset ~100-1000 😩



Mayhew et. al. (2022)

What are foundation models?

- Neural networks trained on large amounts of unlabeled data.
- Models are then transferred to downstream tasks.
- Good at generalizing to small datasets.



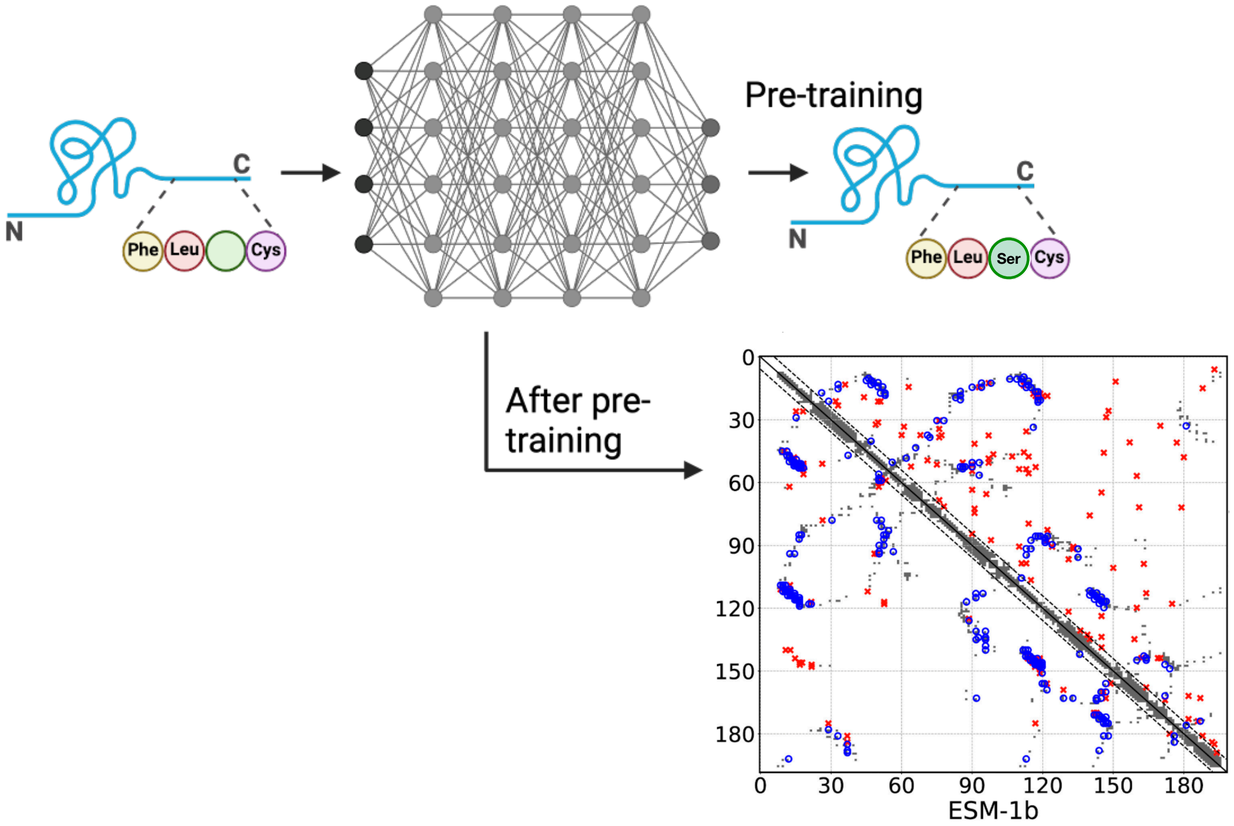
Gidaris et. al. (2018)

Foundation models in protein biology

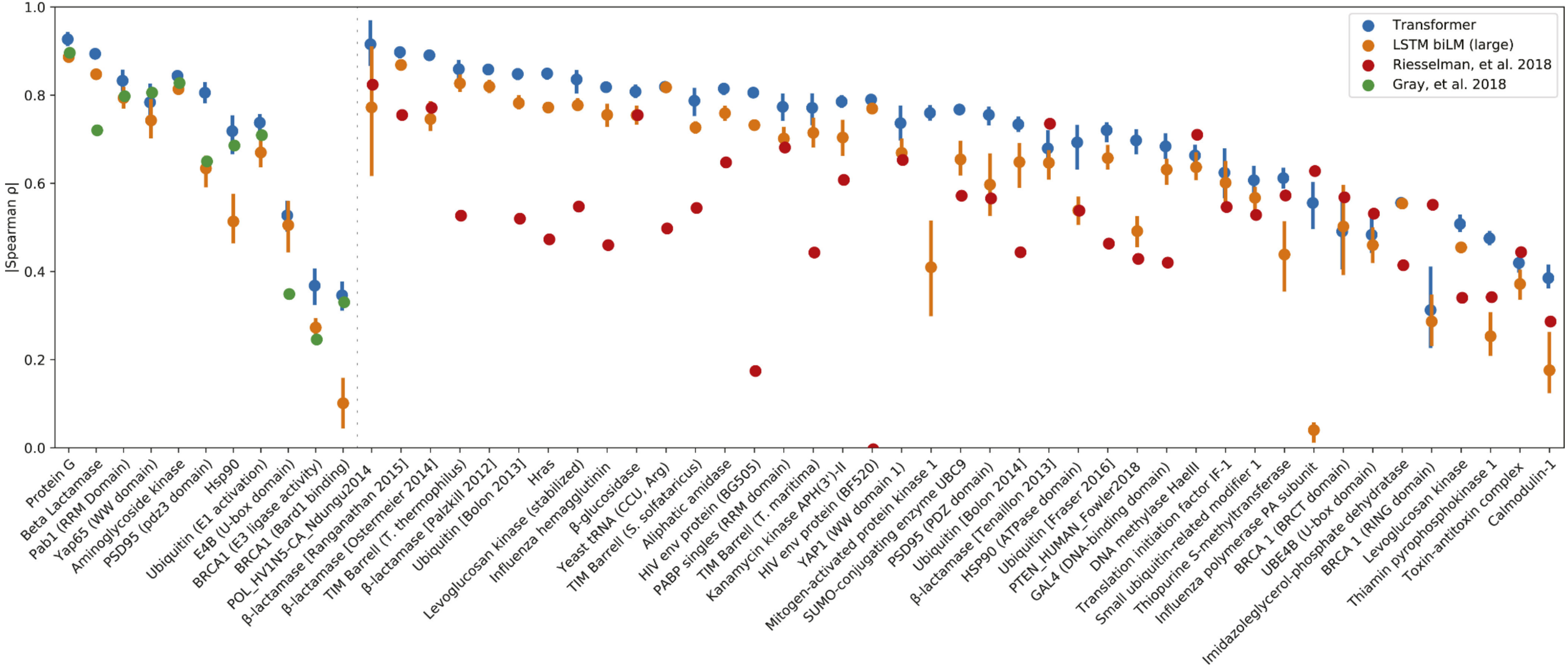
RESEARCH ARTICLE | BIOLOGICAL SCIENCES |

Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences

Alexander Rives, Joshua Meier, Tom Sercu, +7, and Rob Fergus Authors Info & Affiliations



Success of ESM

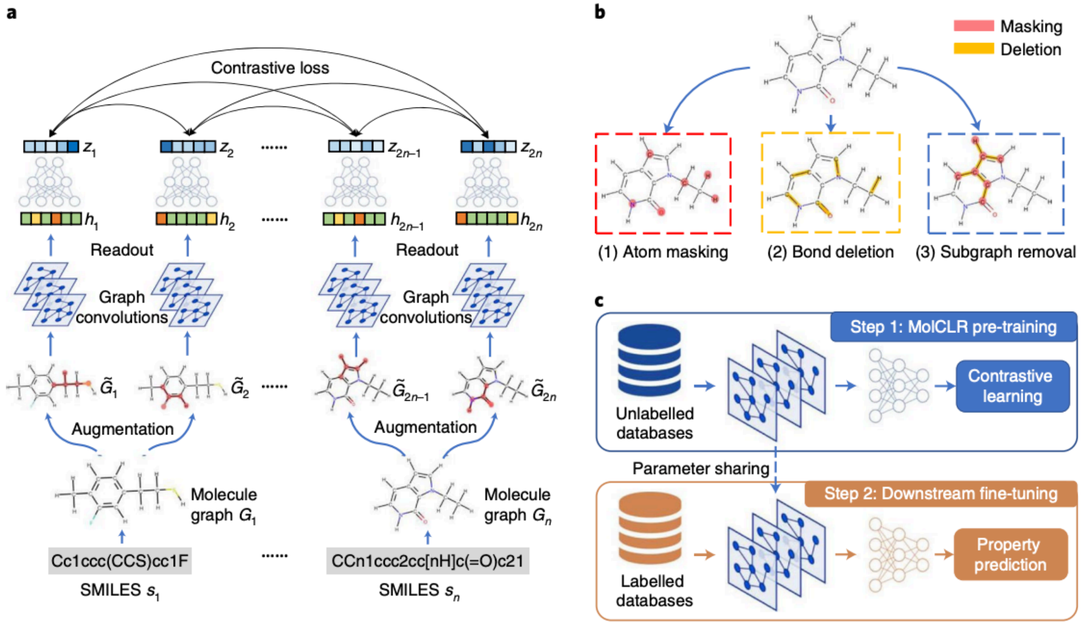


Rives (2021)

Foundation models in chemistry

Molecular contrastive learning of representations via graph neural networks

Yuyang Wang[1,2], Jianren Wang[3], Zhonglin Cao[1] and Amir Barati Farimani[1,2,4] ✉



ChemBERTa-2: Towards Chemical Foundation Models

Walid Ahmad*
Reverie Labs
walid@reverielabs.com

Elana Simon*
Reverie Labs
elana@reverielabs.com

Seyone Chithrananda
UC Berkeley
seyonec@berkeley.edu

Gabriel Grand
Reverie Labs & MIT CSAIL
gg@mit.edu

Bharath Ramsundar
Deep Forest Sciences
bharath@deepforestsci.com

MLM: Masked Language Model

C1CC1(C(=O)NC2=CC=C(C=C2)OC3=C4C=C(NC4=NC=C3)C(=O)NCCN5CCOCC5)C(=O)NC6=CC=C(C=C6)F

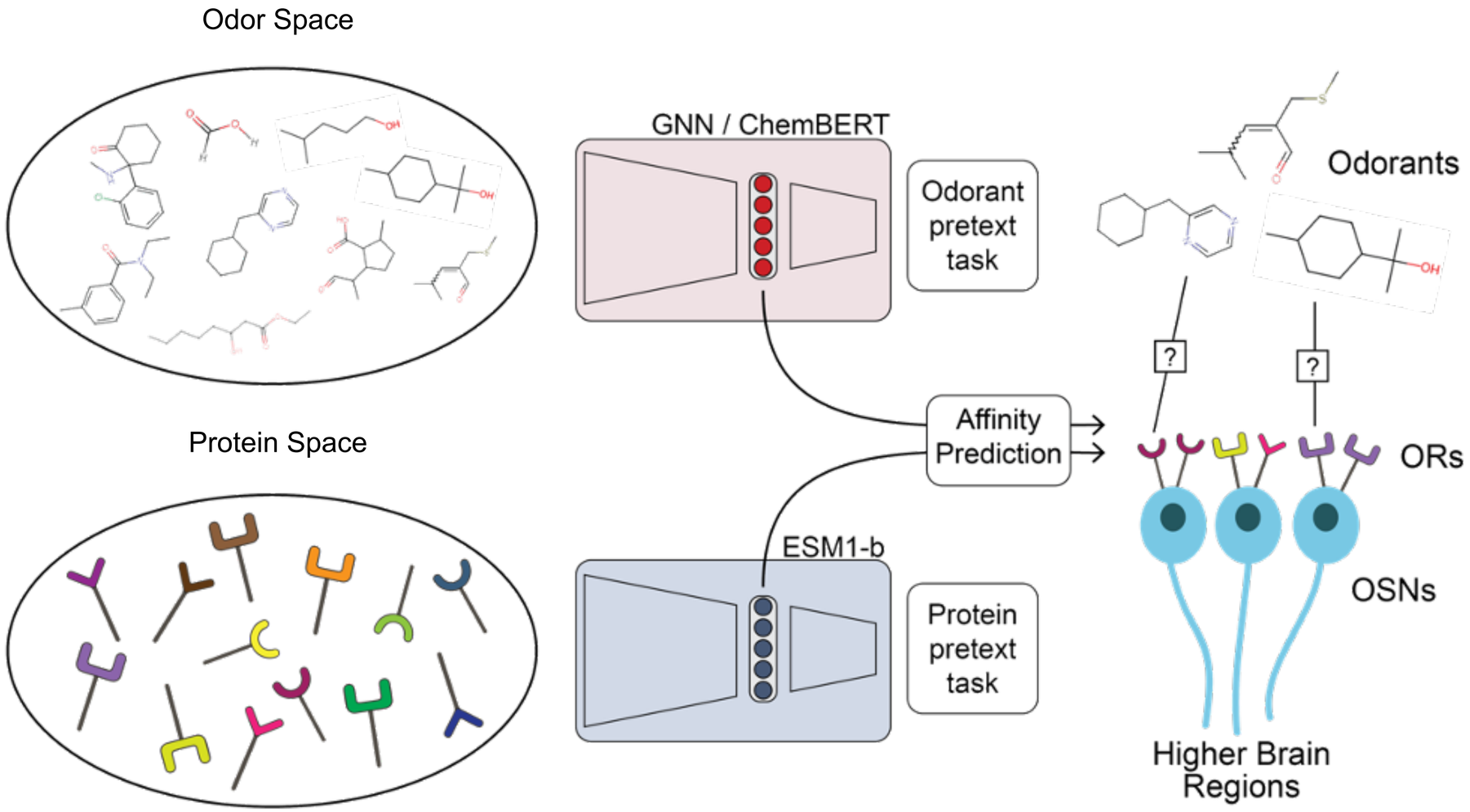
C1CC1(C(=O)NC2=CC=C(C=C2)OC3=C4C=C(NC4=NC=C3)C(=O)NCCN5CCOCC5)C(=O)NC6=CC=C(C=C6)F

MTR: Multi-Task Regression

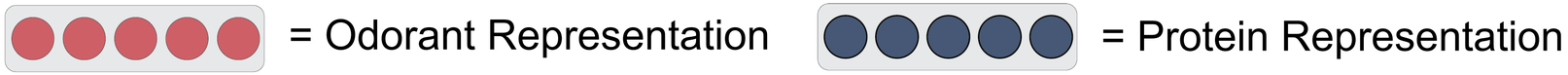
C1CC1(C(=O)NC2=CC=C(C=C2)OC3=C4C=C(NC4=NC=C3)C(=O)NCCN5CCOCC5)C(=O)NC6=CC=C(C=C6)F

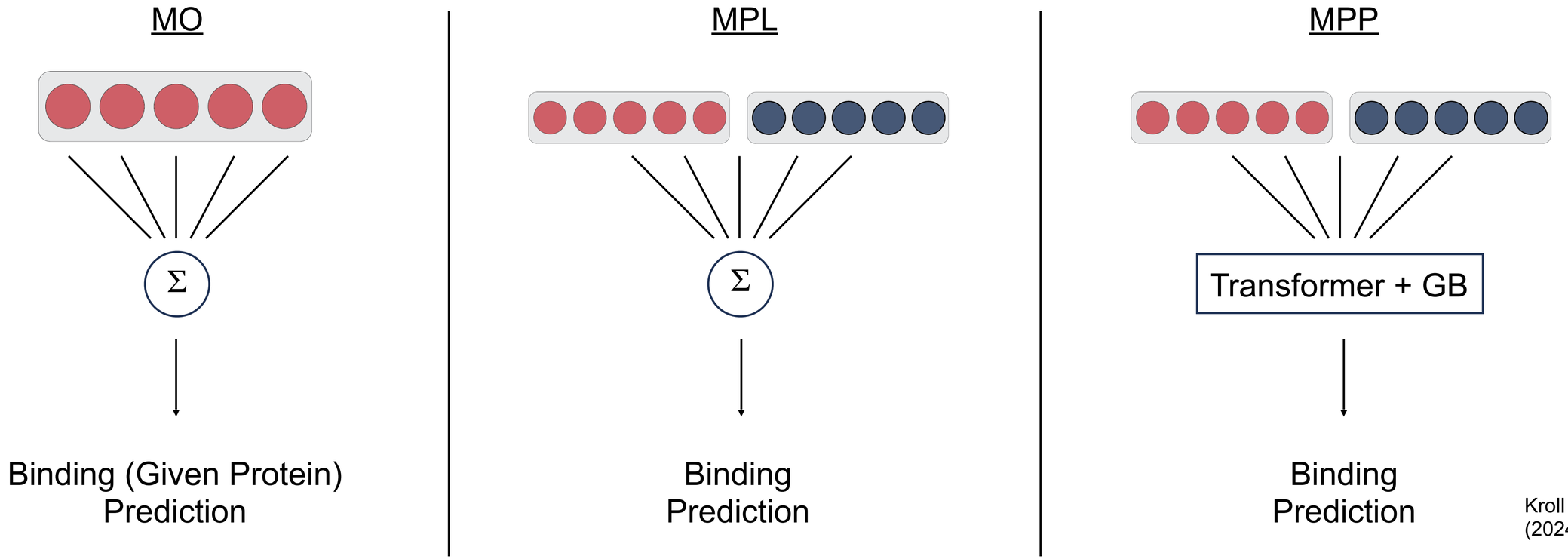
MW, TPSE, FSP3C, #ROTBONDS, etc ...

P2: Multimodal foundation models for small datasets



Affinity prediction model variants





Kroll et. al. (2024)

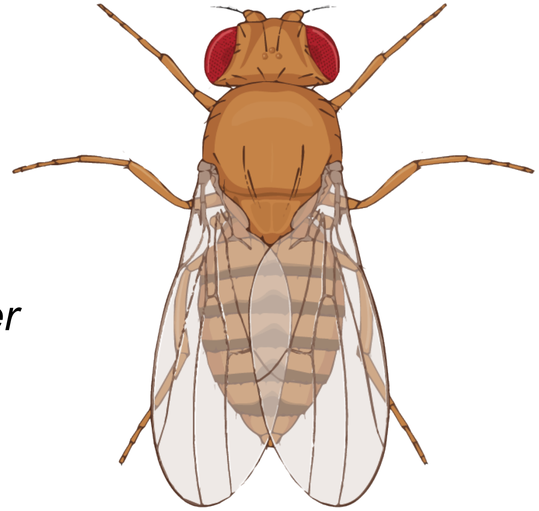
Can foundation models improve odorant-OR prediction?

- Q1: How well can we predict single OR response with chemical embeddings?
- Q2: How does adding protein information improve prediction?
- Q3: Are different chemical embeddings better than others?

Datasets

Hallem et. al. (2006)

Drosophila melanogaster



110 odorants
2640 pair responses
EC50 (spikes/s)

Lalis et. al. (2024)

Homo sapiens

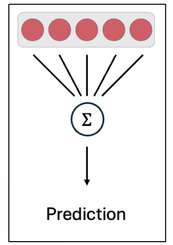
Mus musculus

474 odorants
5835 pair responses
EC50 (binary)

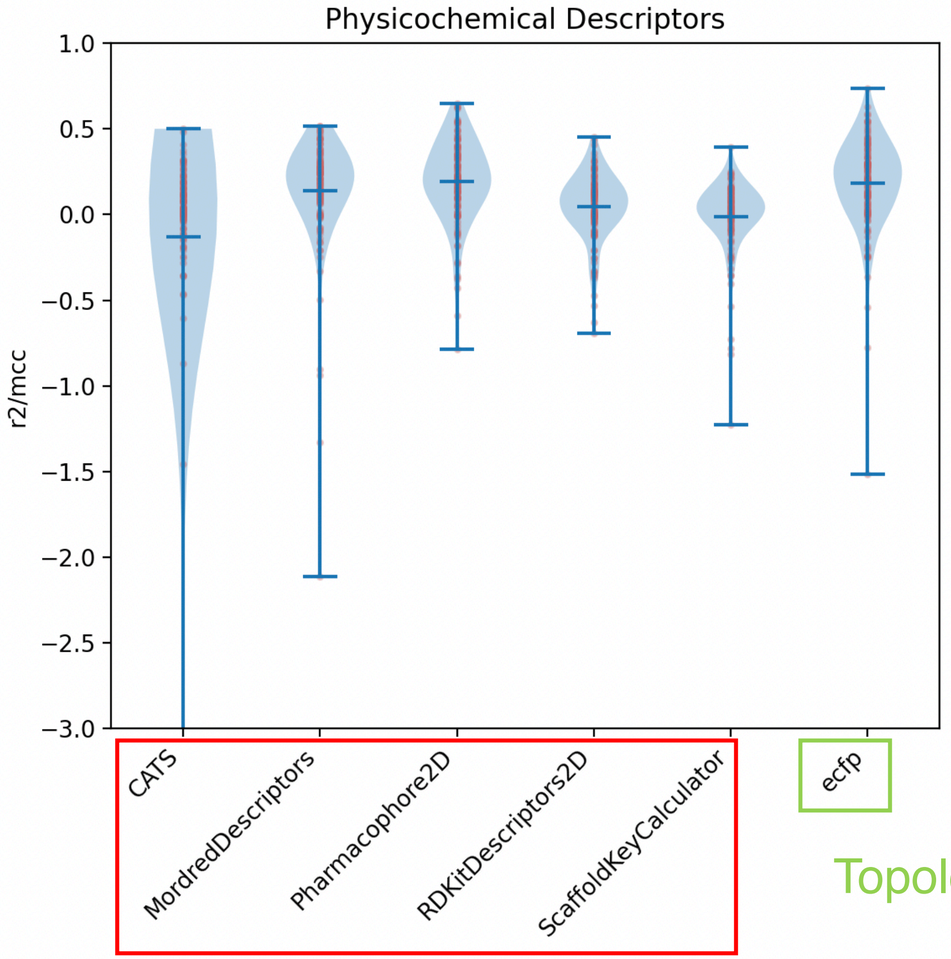
Can foundation models improve odorant-OR prediction?

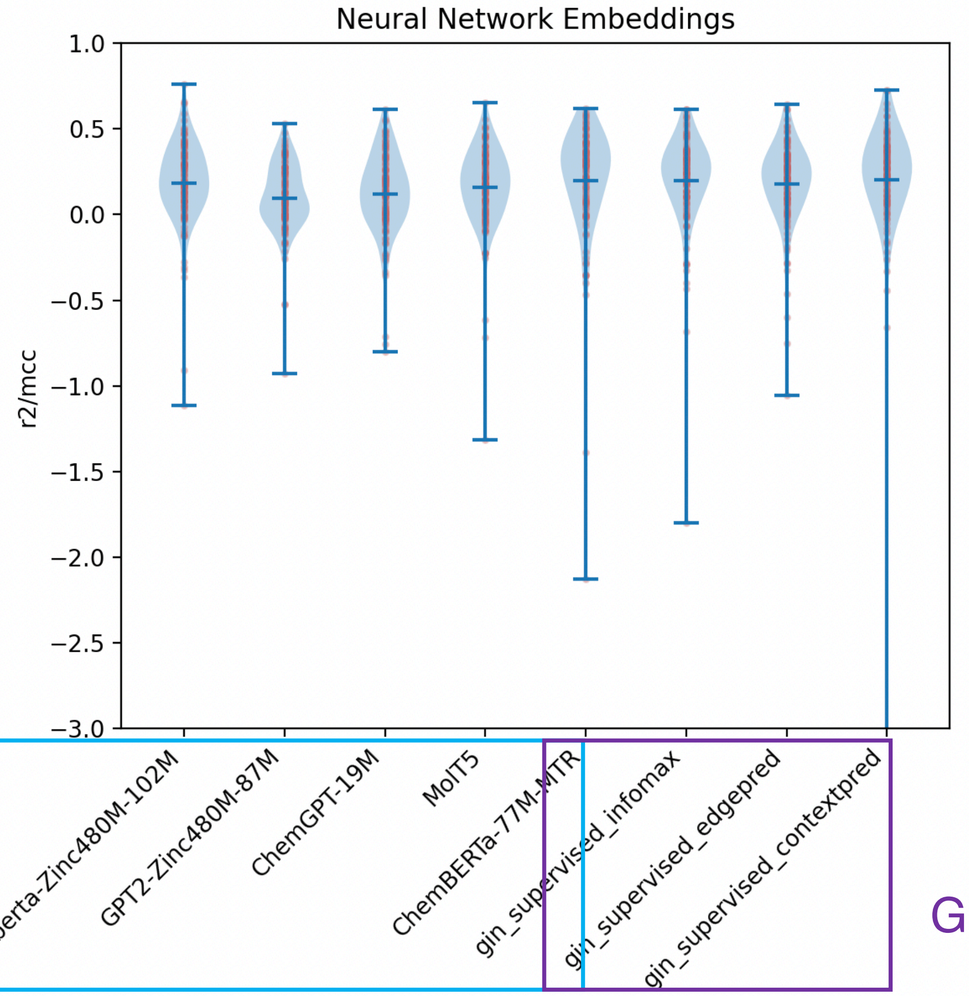
- Q1: How well can we predict single OR response with chemical embeddings?
- Q2: How does adding protein information improve prediction?
- Q3: Are different chemical embeddings better than others?

Q1: Chemistry alone is not enough



Hallem – *Drosophila*





Physicochemical

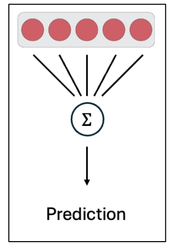
Topological

Transformer

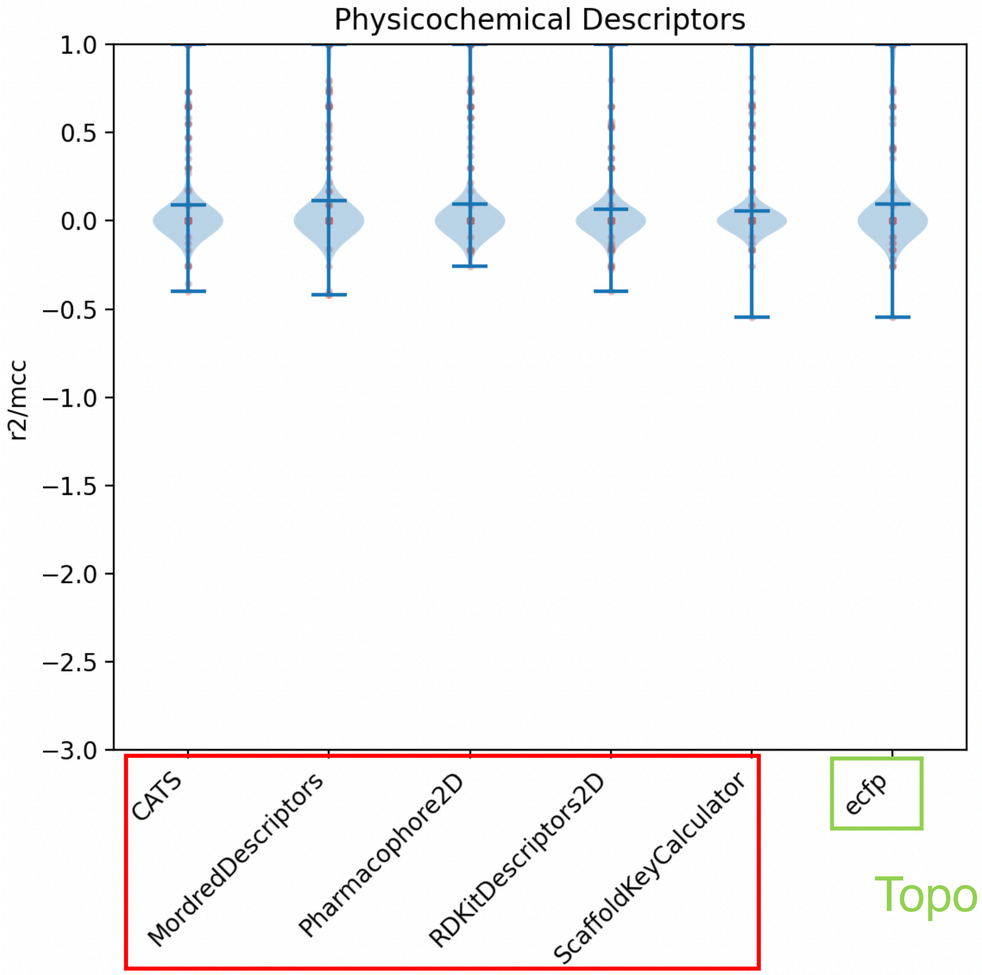
GNNs

McConachie et. al, submitted

Q1: Chemistry alone is not enough

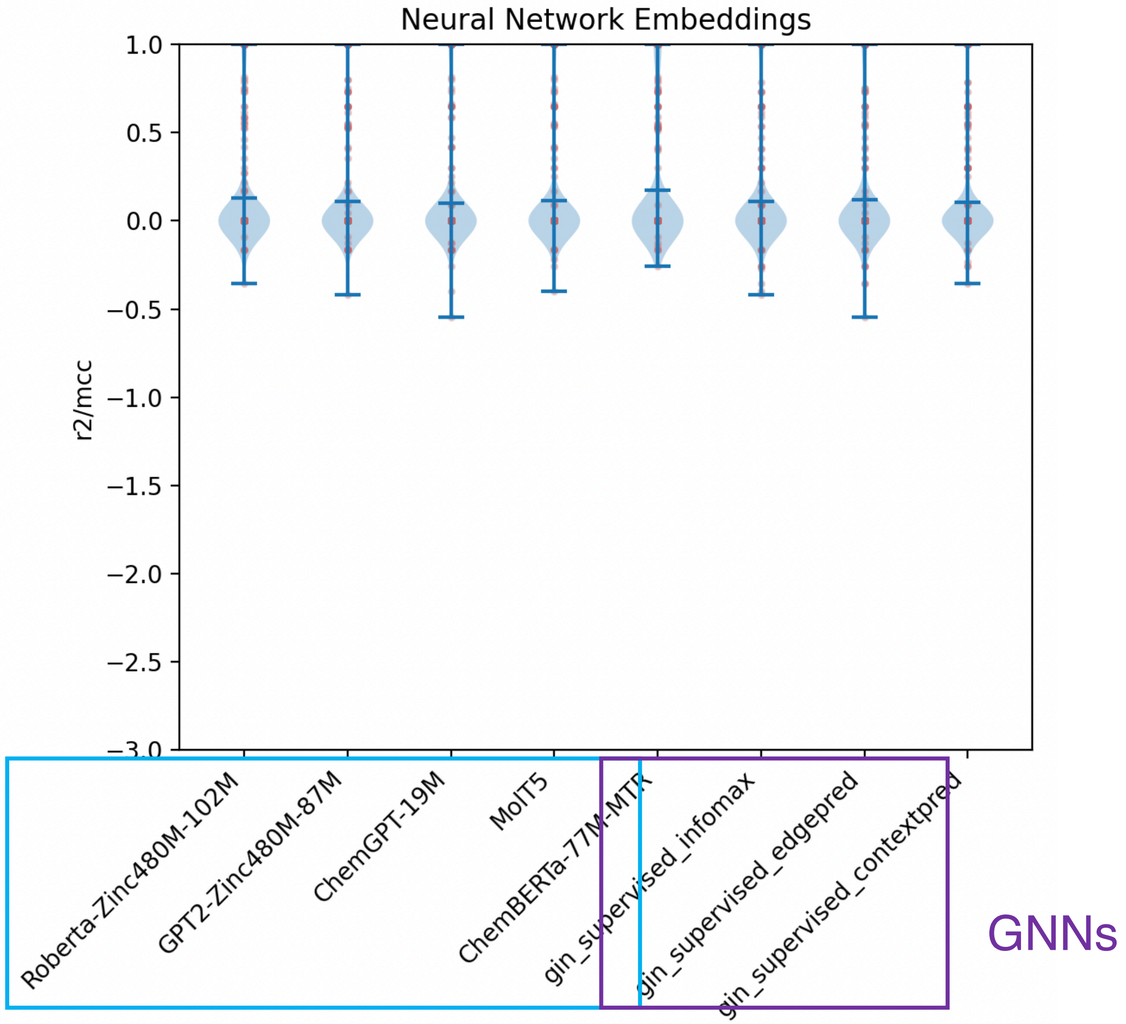


Lalis – Mammalian



Physicochemical

Topological



Transformer

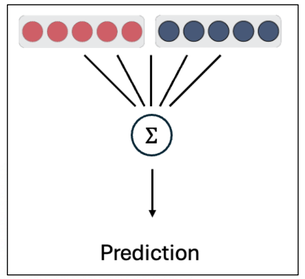
GNNs

McConachie et. al, submitted

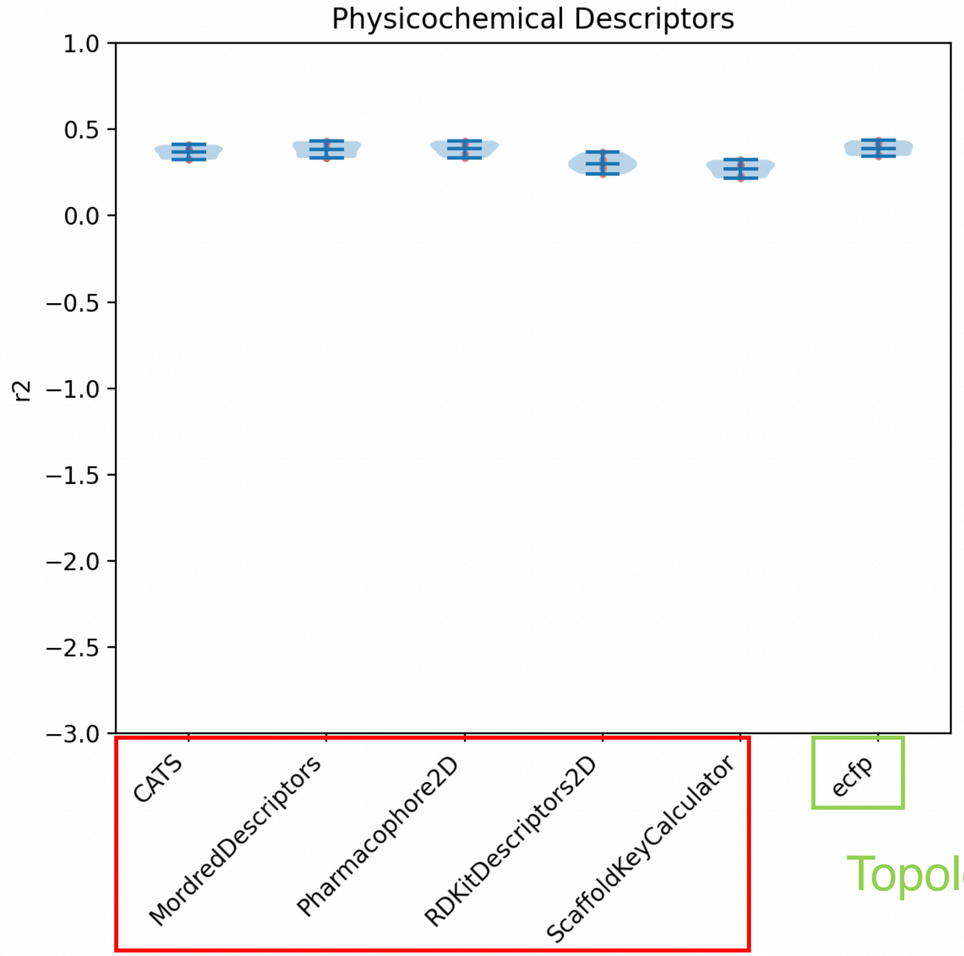
Can foundation models improve odorant-OR prediction?

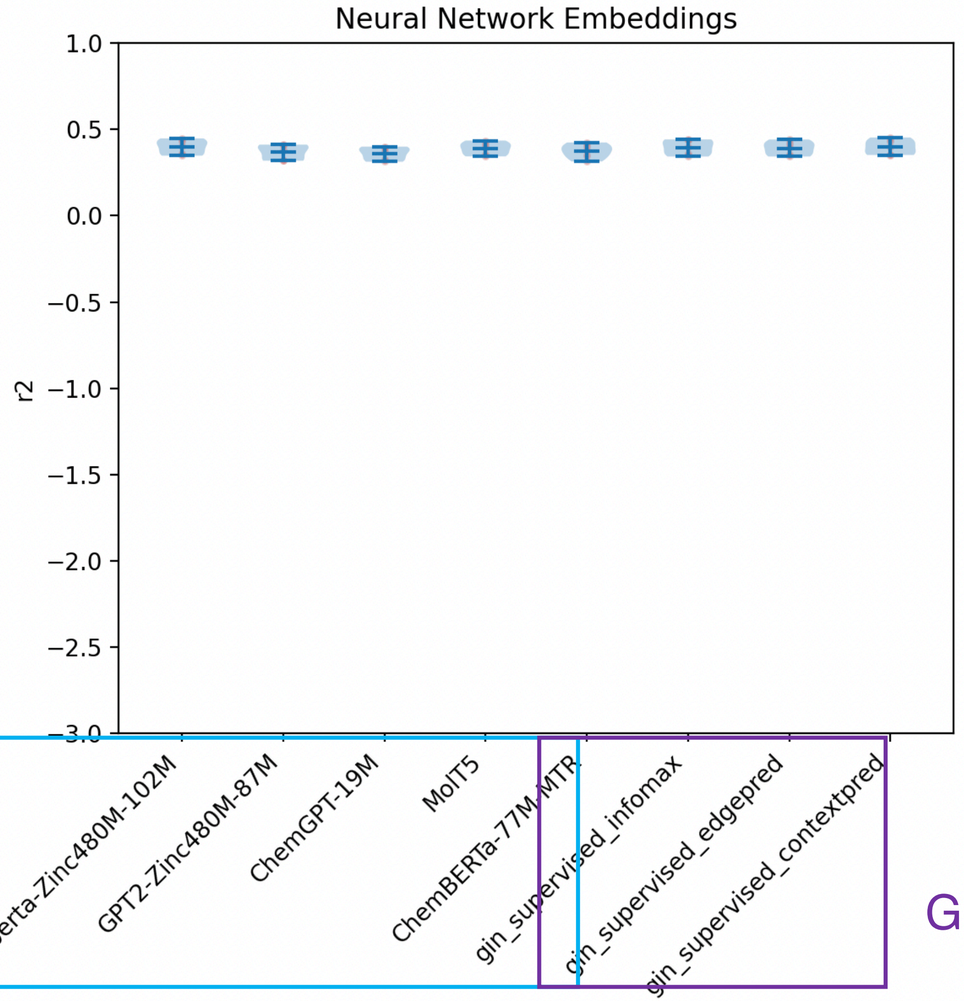
- Q1: How well can we predict single OR response with chemical embeddings?
- Q2: How does adding protein information improve prediction?
- Q3: Are different chemical embeddings better than others?

Q2: Protein embeddings improve prediction



Hallem – *Drosophila*





Physicochemical

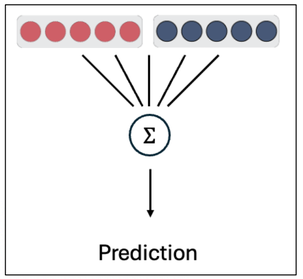
Topological

Transformer

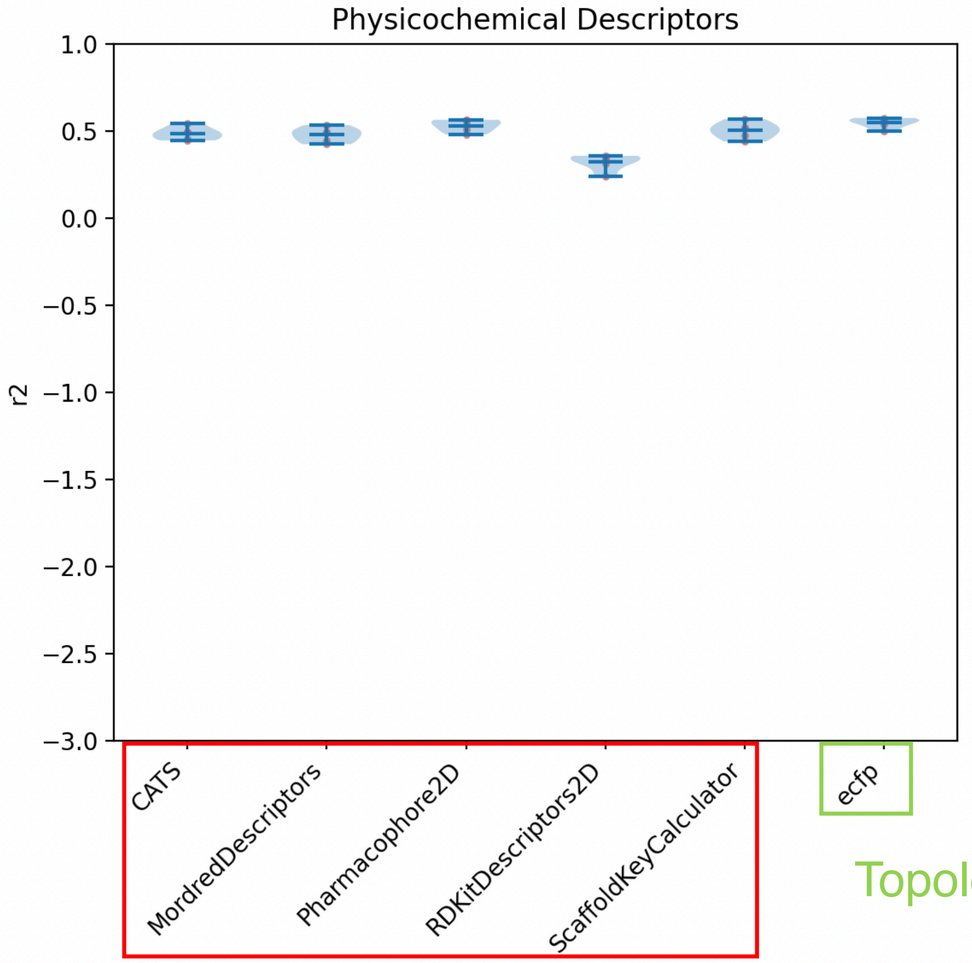
GNNs

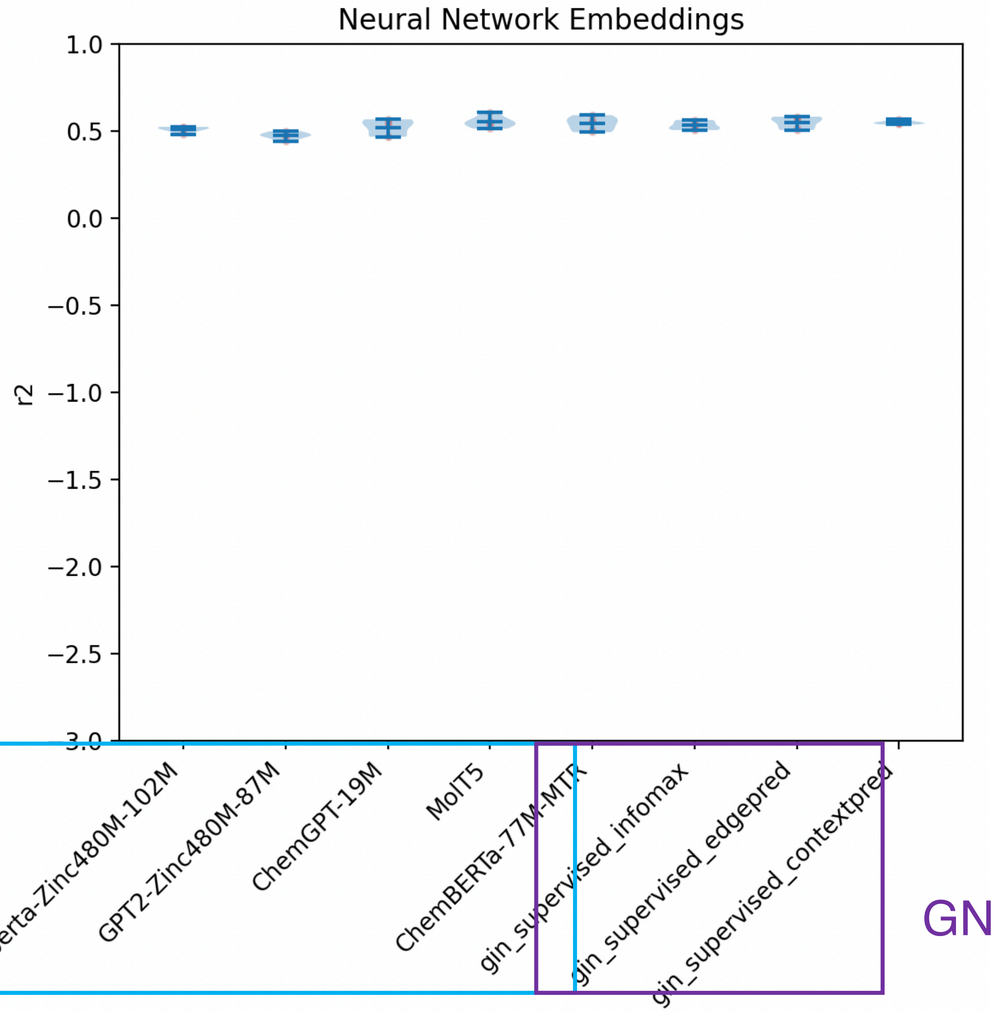
McConachie et. al, submitted

Q2: Protein embeddings improve prediction



Lalis – Mammalian





Physicochemical

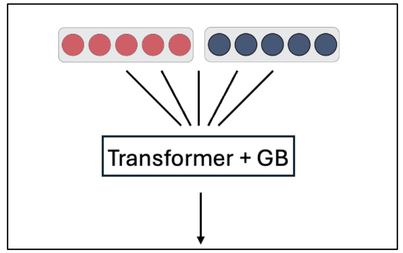
Topological

Transformer

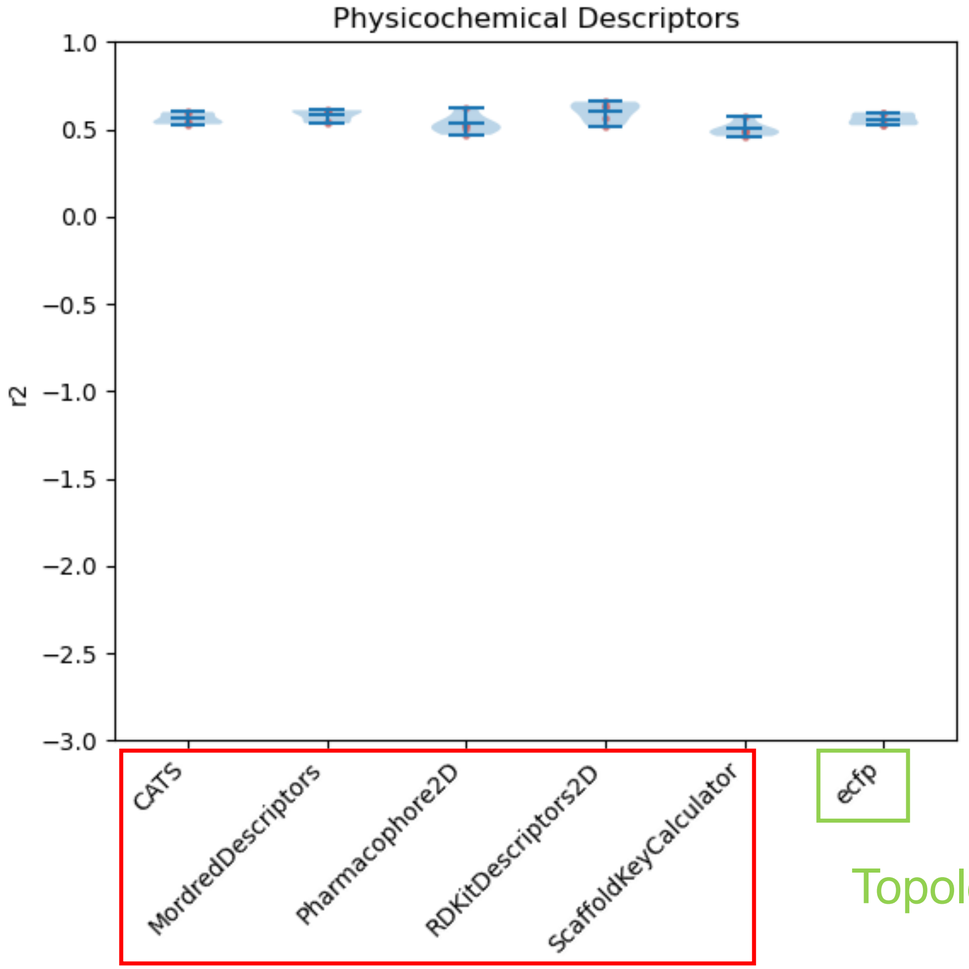
GNNs

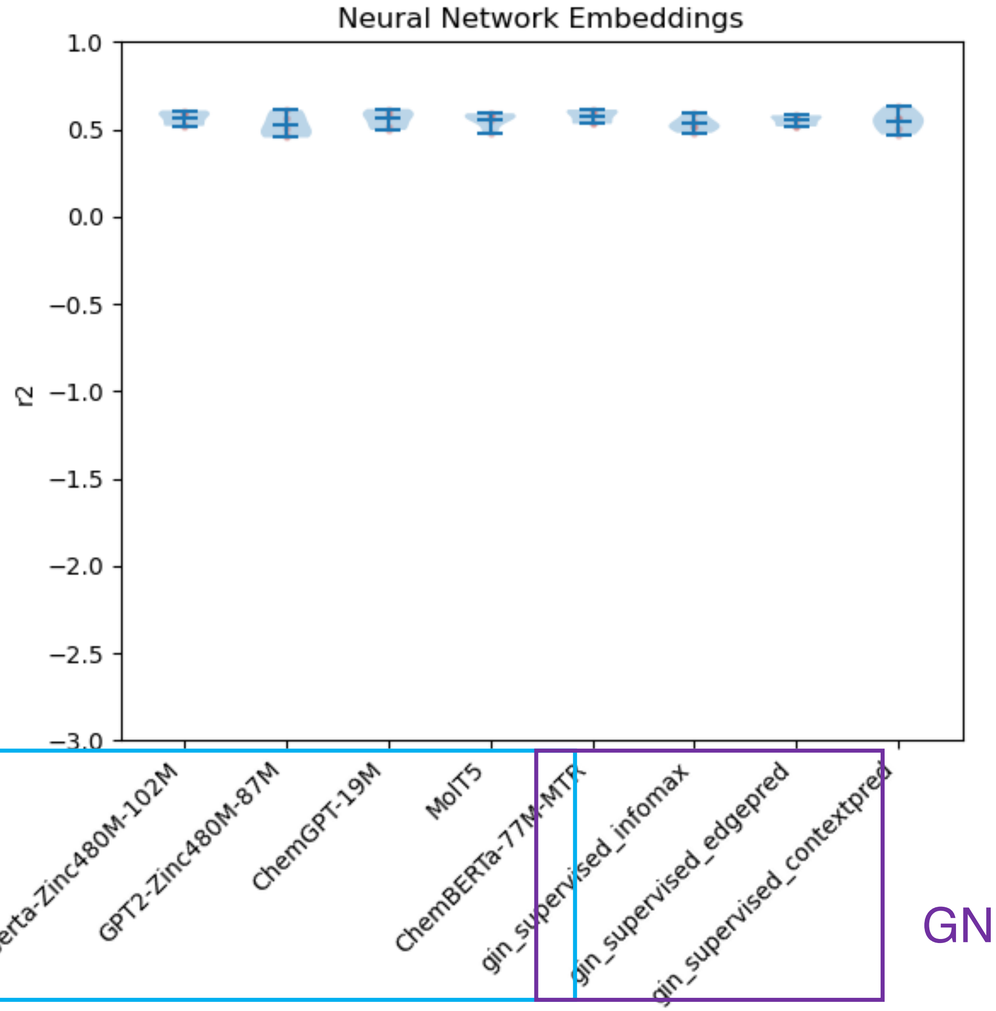
McConachie et. al, submitted

Q2: Nonlinear decoder improves further



Hallem – *Drosophila*





Physicochemical

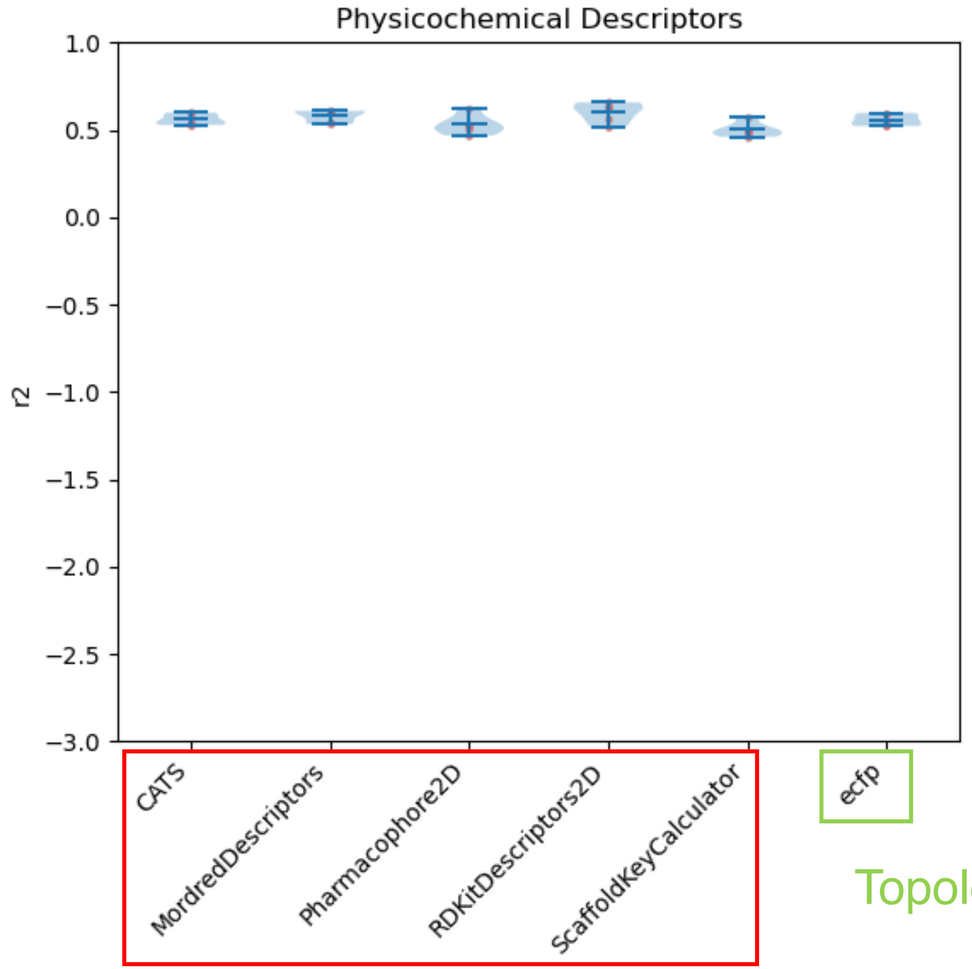
Topological

Transformer

GNNs

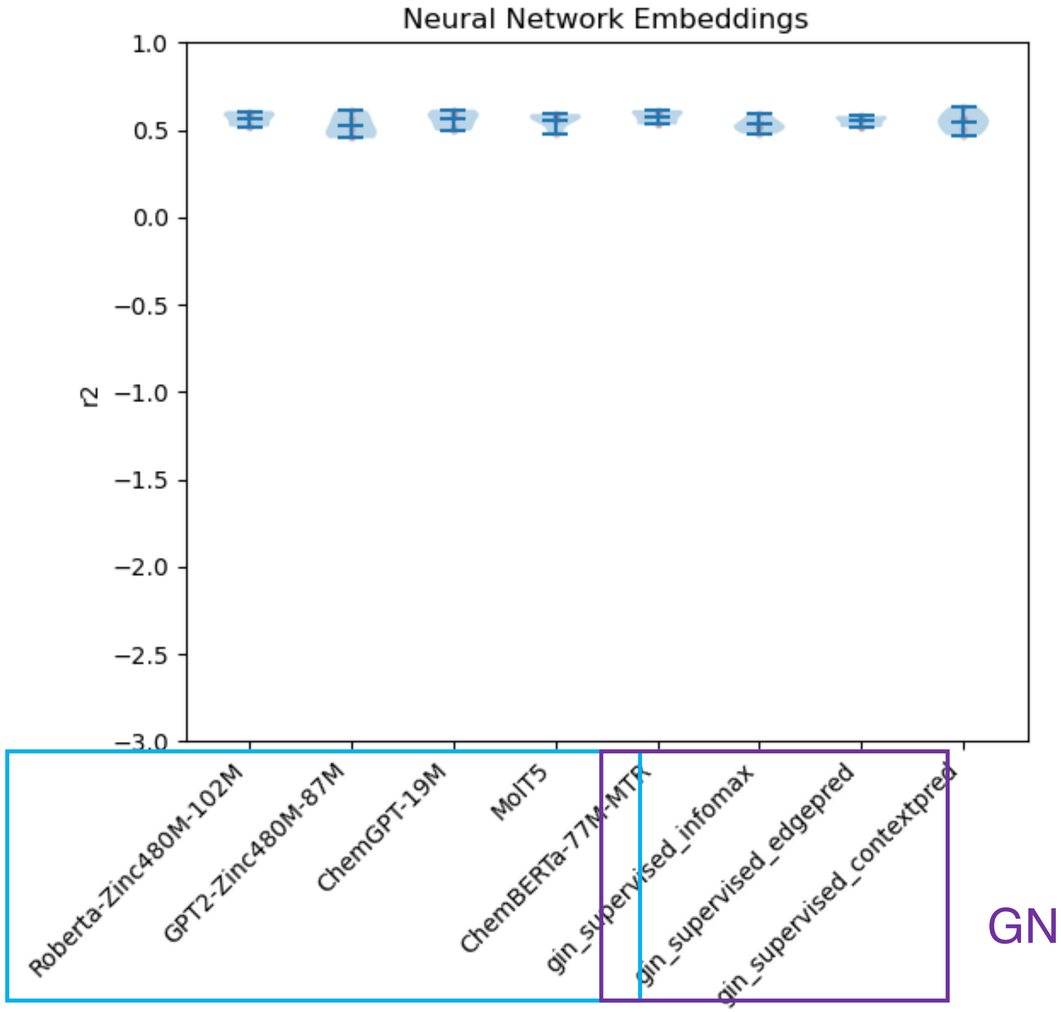
McConachie et. al, submitted

Q3: Chemical foundation models don’t help



Physicochemical

Topological



Transformer

GNNs

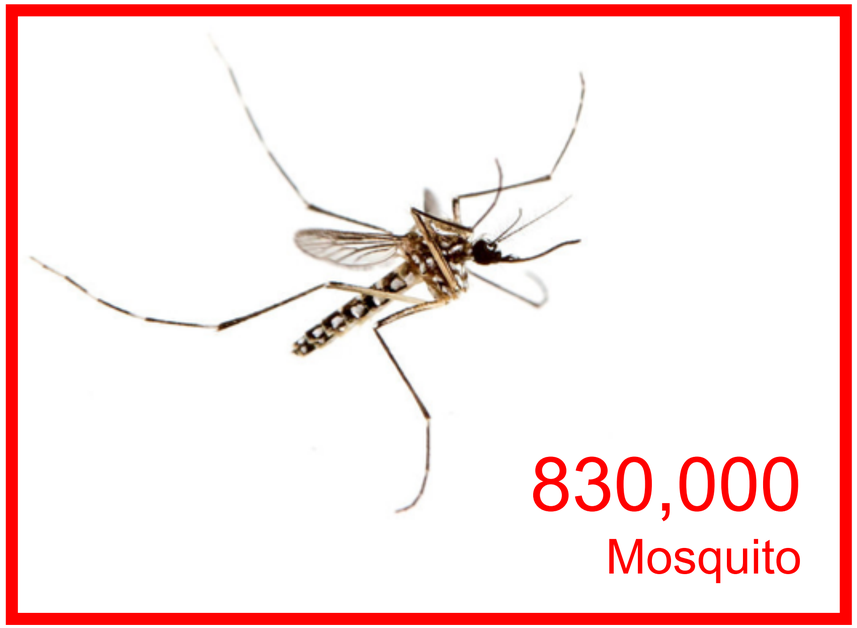
McConachie et. al, submitted

Can foundation models improve odorant-OR prediction?

- Q1: How well can we predict single OR response with chemical embeddings? Not very well.
- Q2: How does adding protein information improve prediction? A lot!
- Q3: Are different chemical embeddings better than others? Not really, in fact they aren't better than physicochemical descriptors.

For pretraining, context is key! Chemical context is wrong….

Mosquitoes kill a lot of people every year



Deaths per year

60,000 Snake

24,000 Sandfly

17,400 Dog

8,000 Kissing bug

4,400 Freshwater snail

3,500 Scorpion

3,500 Tsetse fly

1,600 Tapeworm

500 Hippopotamus

100 Elephant

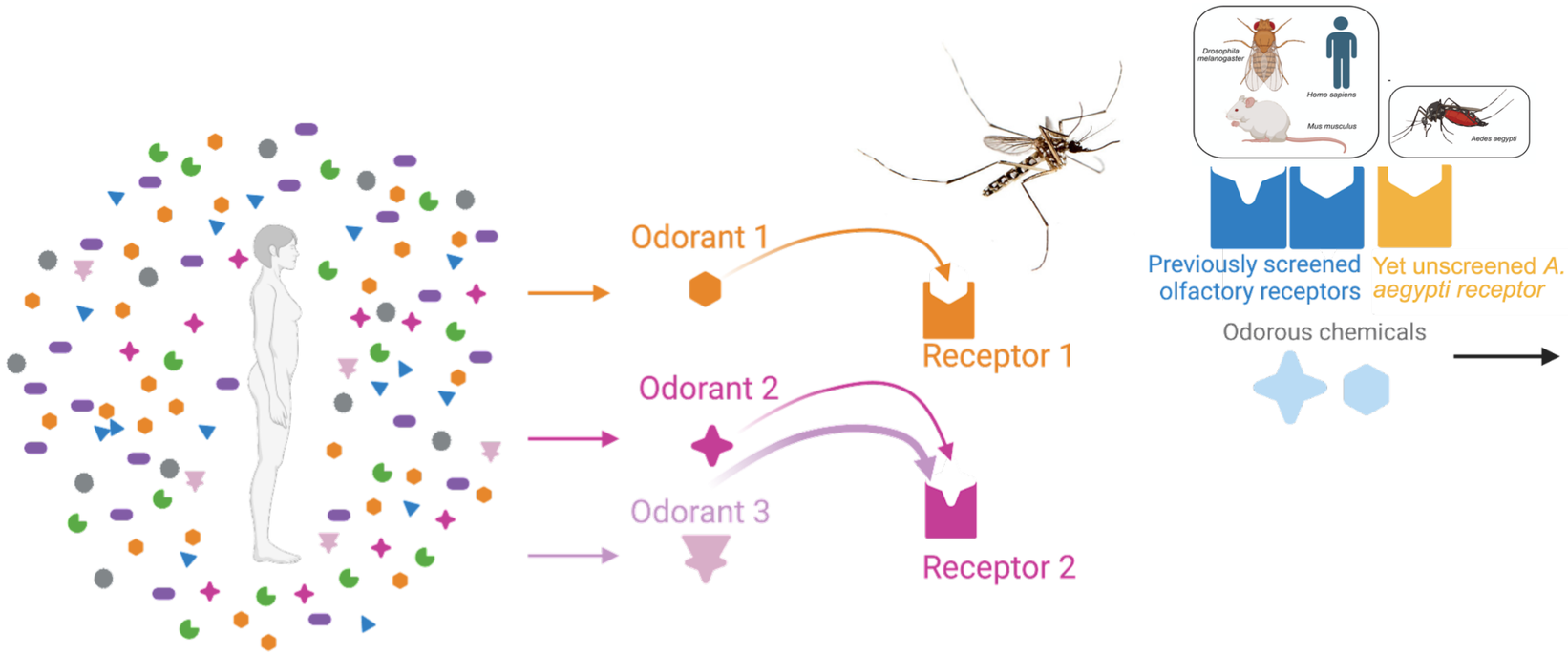
60 Bee

40 Jellyfish

6 Shark

Mosquitoes rely on smell to detect humans

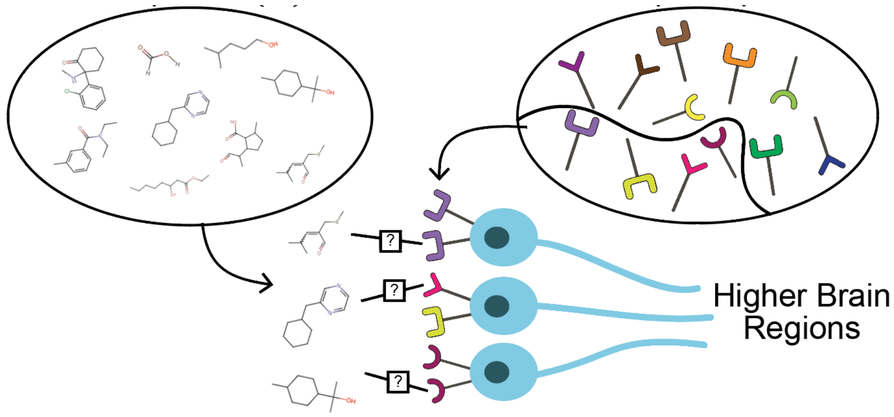
ML guided data collection



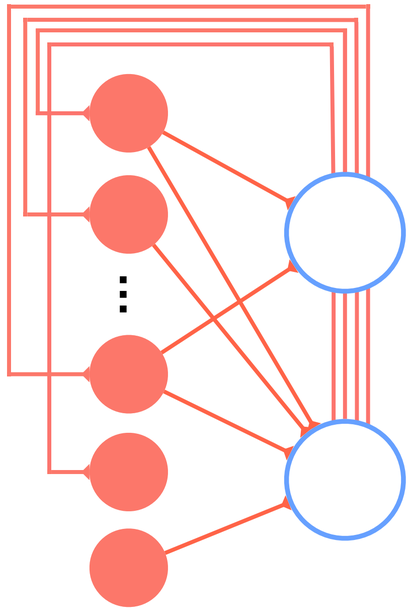
w/ Tori Guerina and Meg Younger

THE DEPAQ LAB @ BU

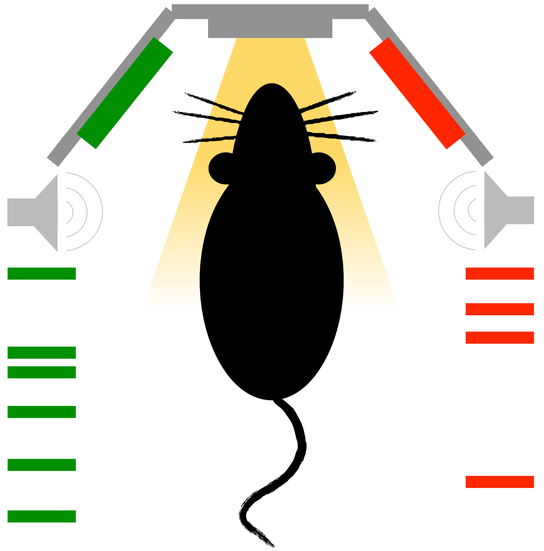
Foundation models for olfaction



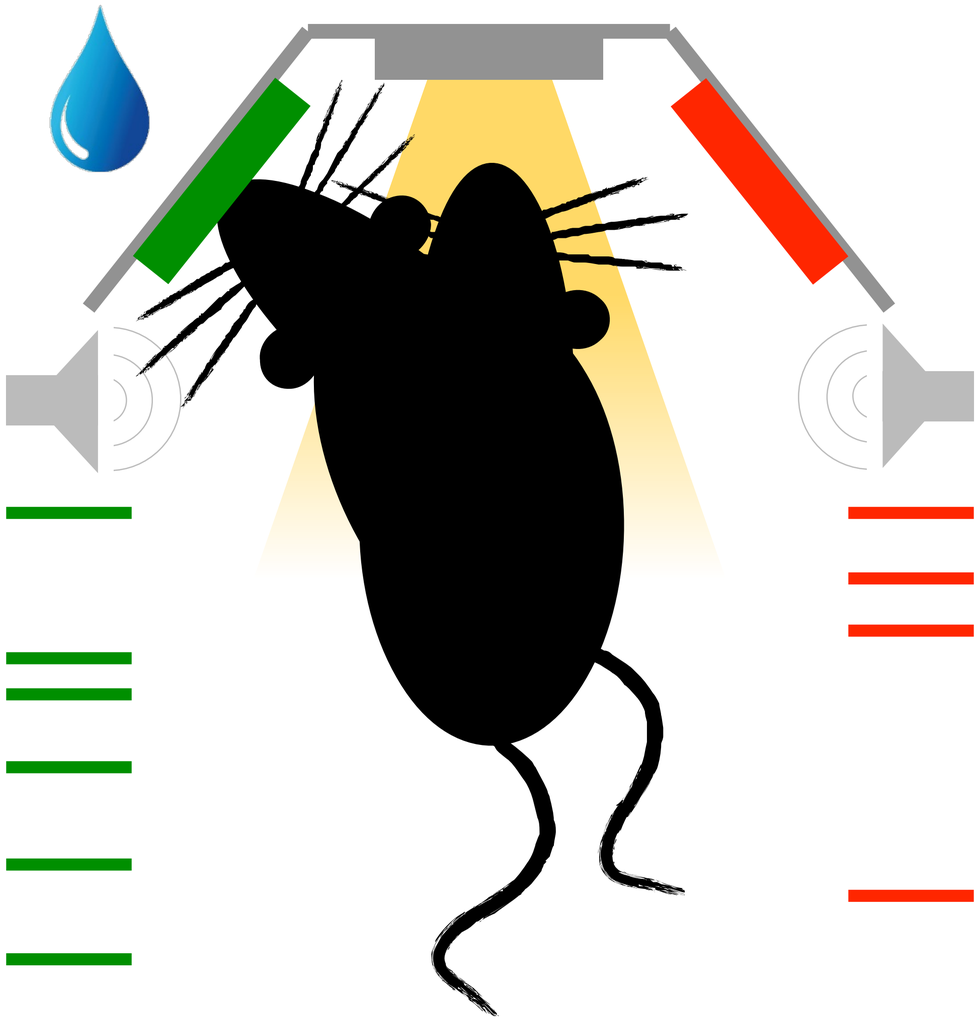
Spike variability from multi-task spiking networks



Hierarchical models of decision-making

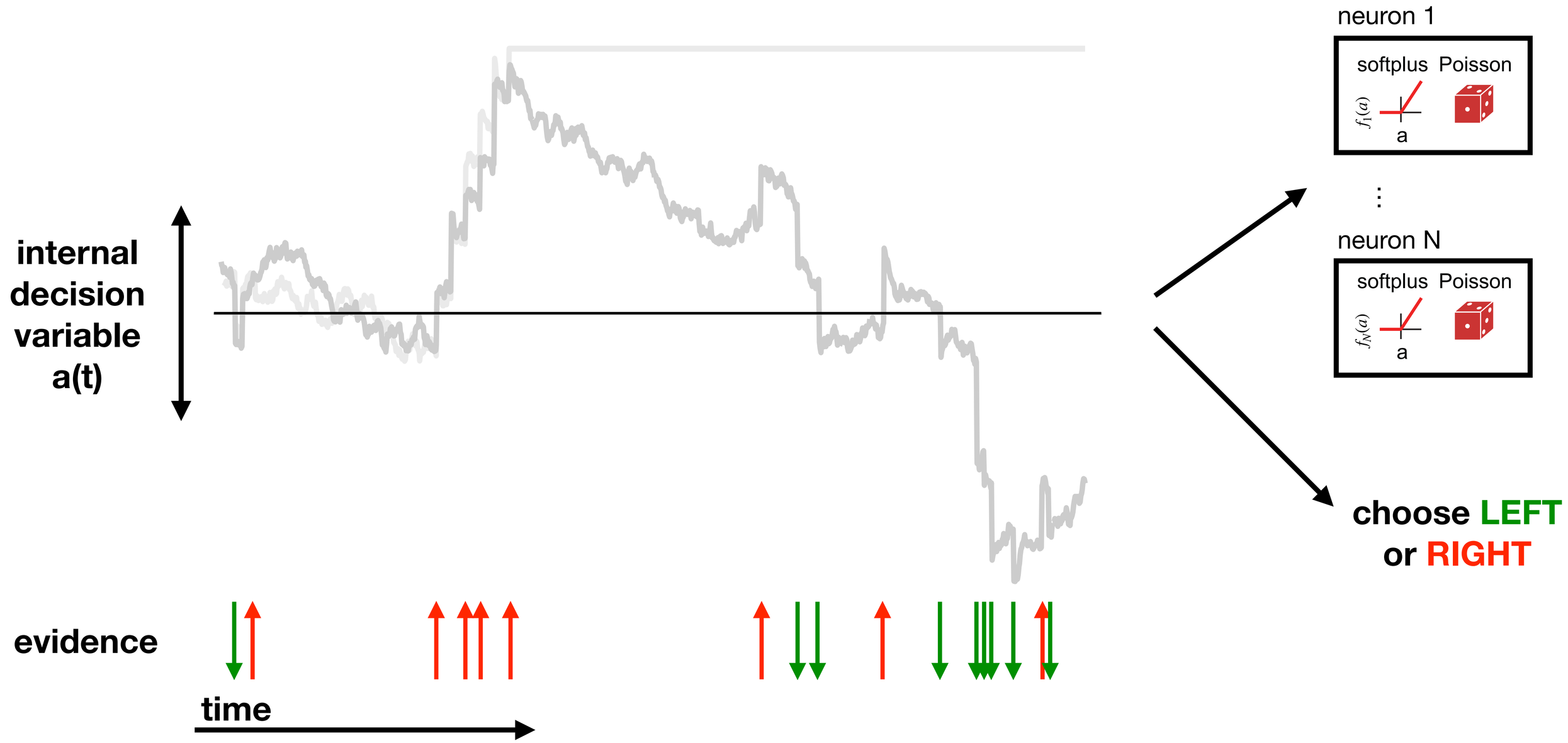


QUESTION: HOW DO WE ACCUMULATE EVIDENCE TOWARDS A DECISION?



Brunton et al. *Science* 2013

MACHINE LEARNING WITH THE DDM

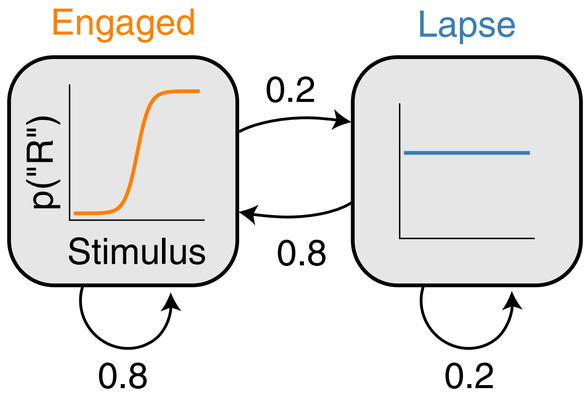


DePasquale, Brody, Pillow. 2024 eLife

STATE-DEPENDENT STRATEGIES

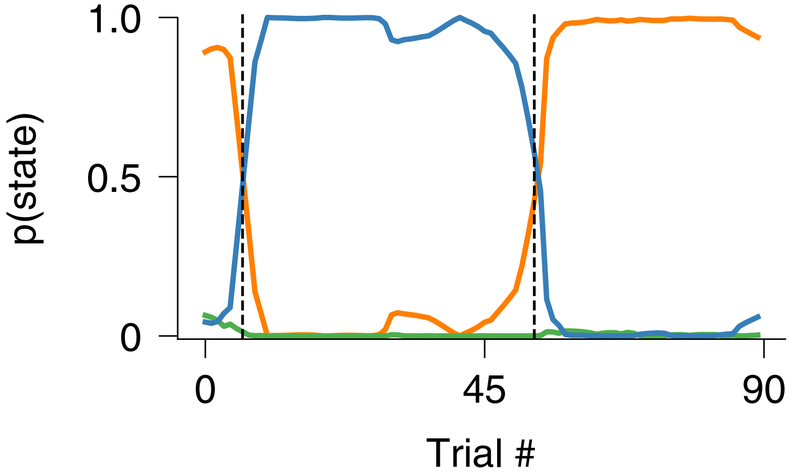
Animals change Decision strategy

Classic lapse model



Example state sequence

Example session 1

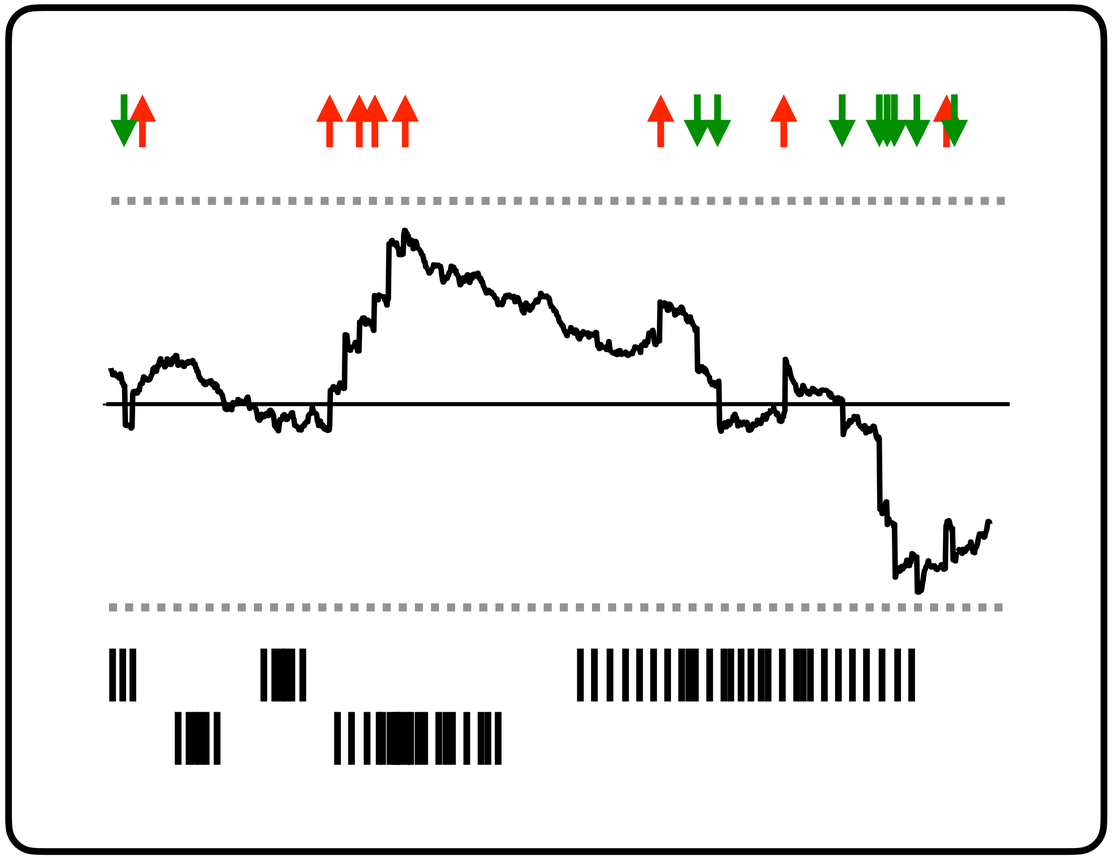


- **What parameters ('strategy') change?**
- **Are there other behavioral signatures of state switches?**

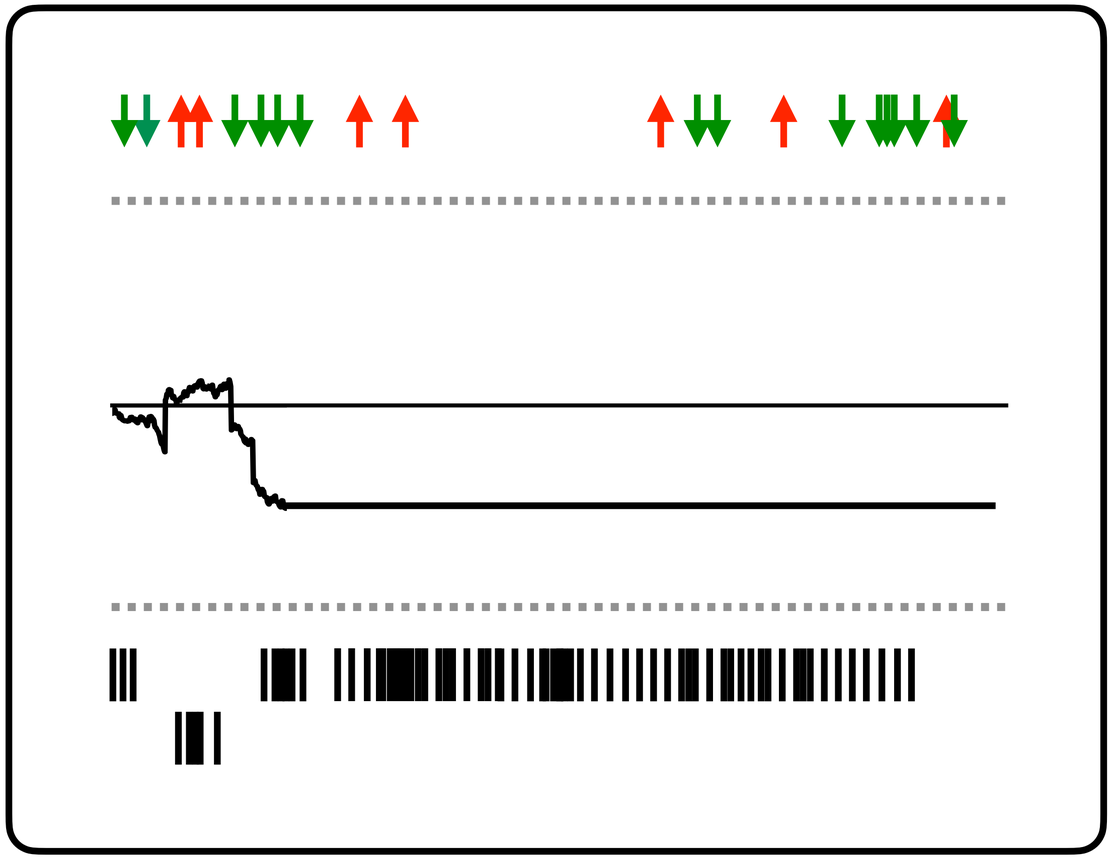
Ashwood et. al, Nat. Neuro, 2022

STATE-DEPENDENT ACCUMULATION MODEL

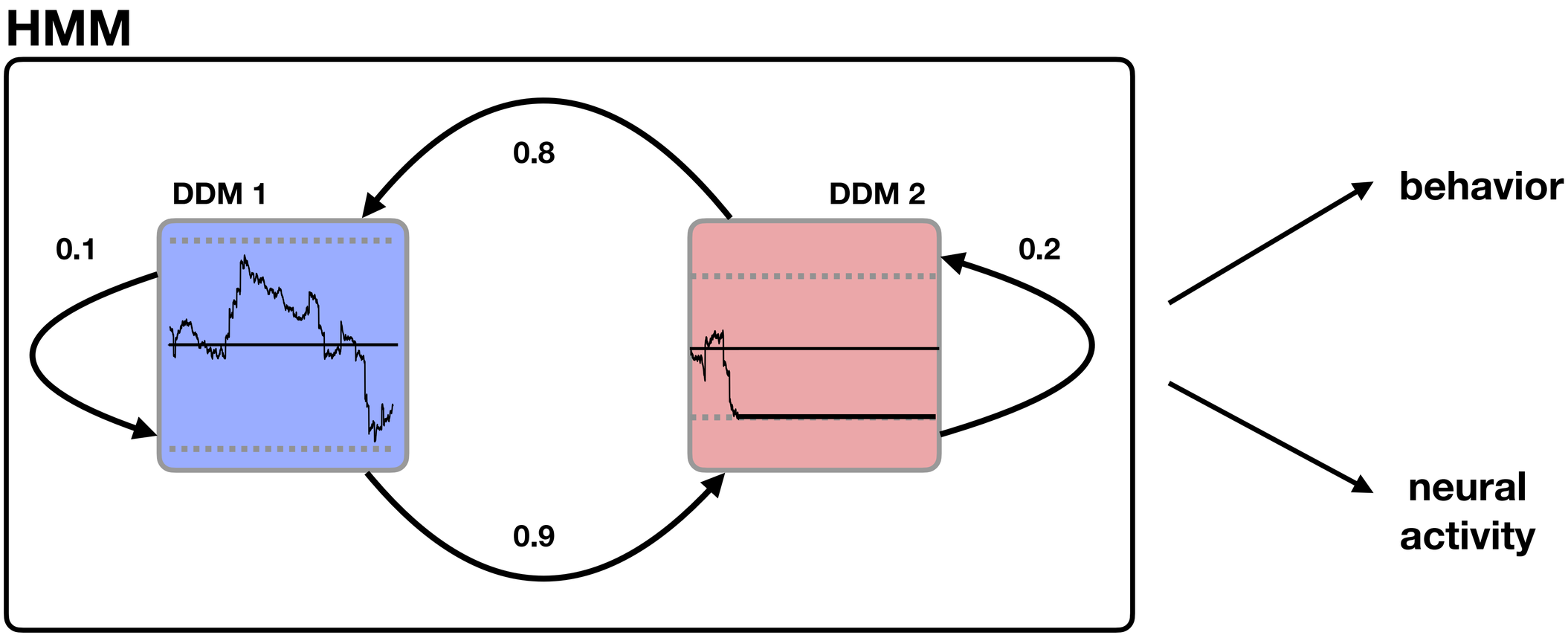
trial 1



trial 2



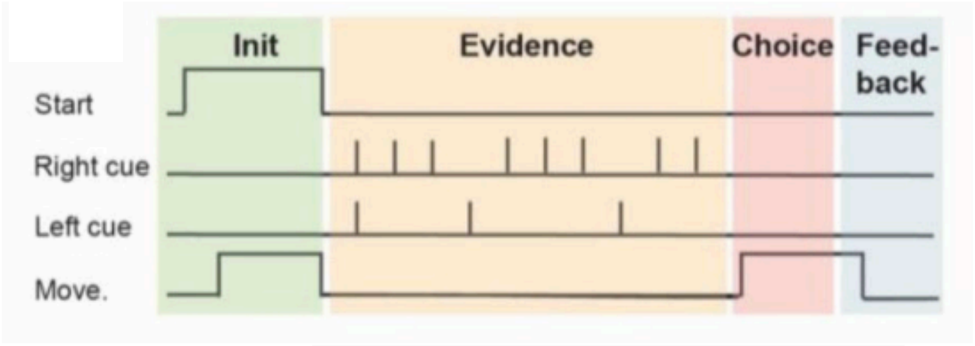
NEW MODEL: HMM-DDM

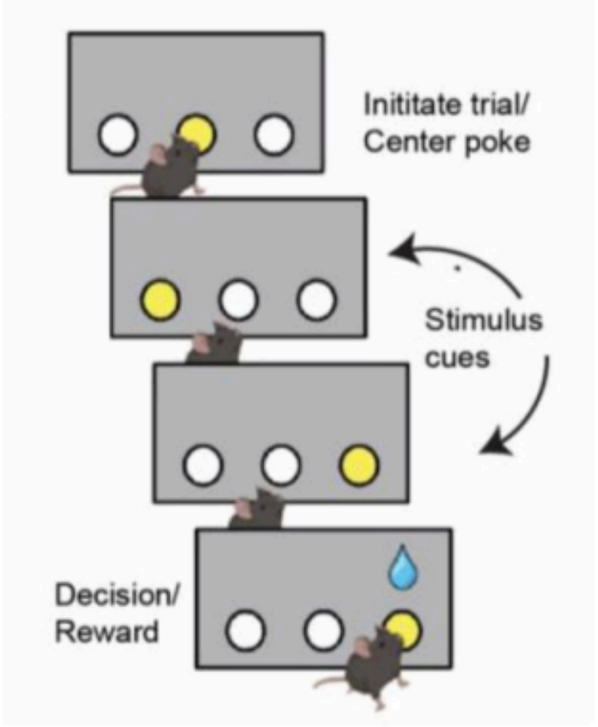


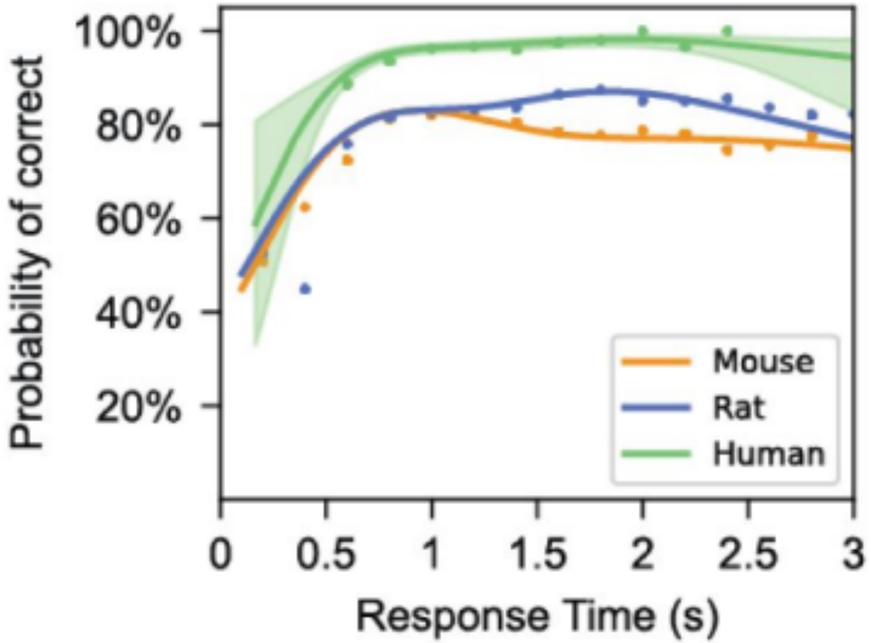
Senne et. al, submitted

w/ Ryan Senne and Ben Scott

VISUAL EVIDENCE ACCUMULATION TASK

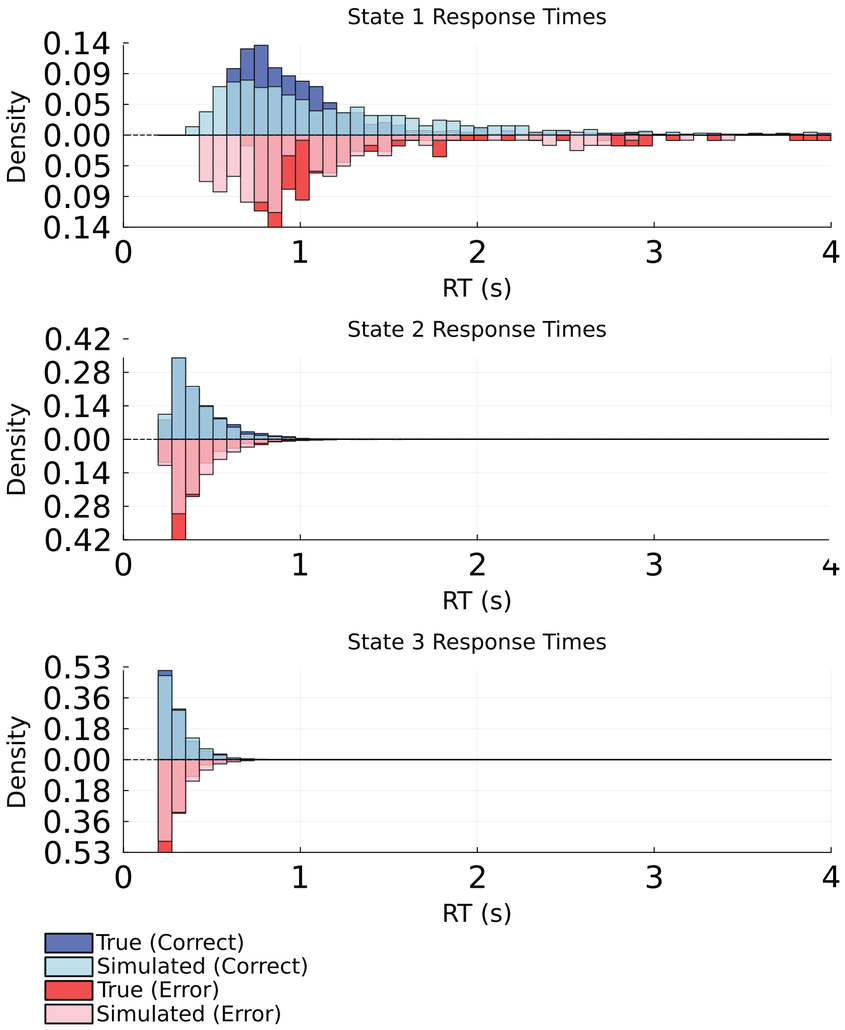


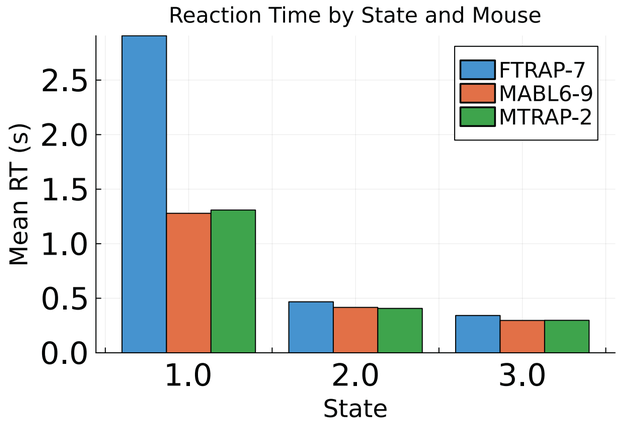


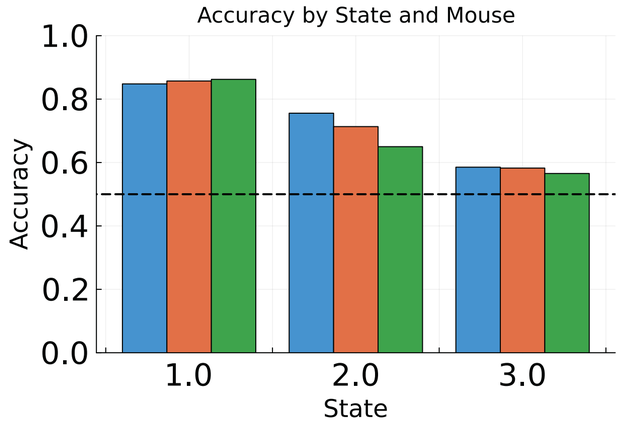


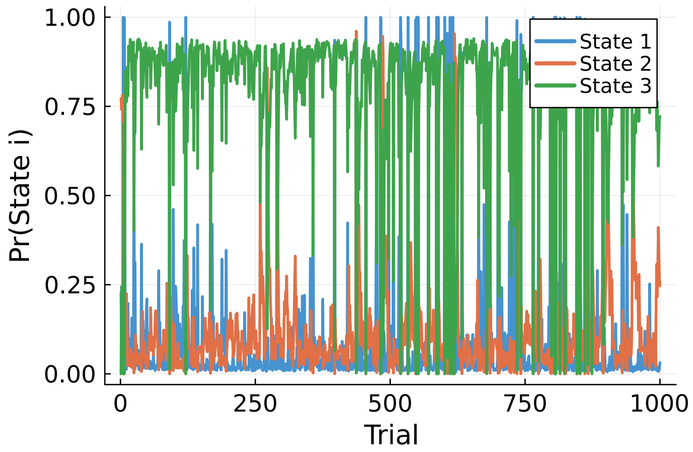
Chakravarty et. al, 2005 eLife

HMM-DDM REVEALS CHANGING SPEED-ACCURACY TRADEOFF









Senne et. al, submitted

w/ Ryan Senne and Ben Scott

StateSpaceDynamics.jl

- Two main models: hidden Markov model (HMM) and linear dynamical system (LDS)
- HMMs are used for hierarchical SSMs but also support other popular models (GLM-HMM)
- LDS models use direct maximization of log joint and global Laplace approximation (Paninski et al. 2009)

README GPL-3.0 license

StateSpaceDynamics.jl: A Julia package for probabilistic state space models (SSMs)

SSM-CI passing | codecov 94% | tested with Aqua.jl | code style blue

Description

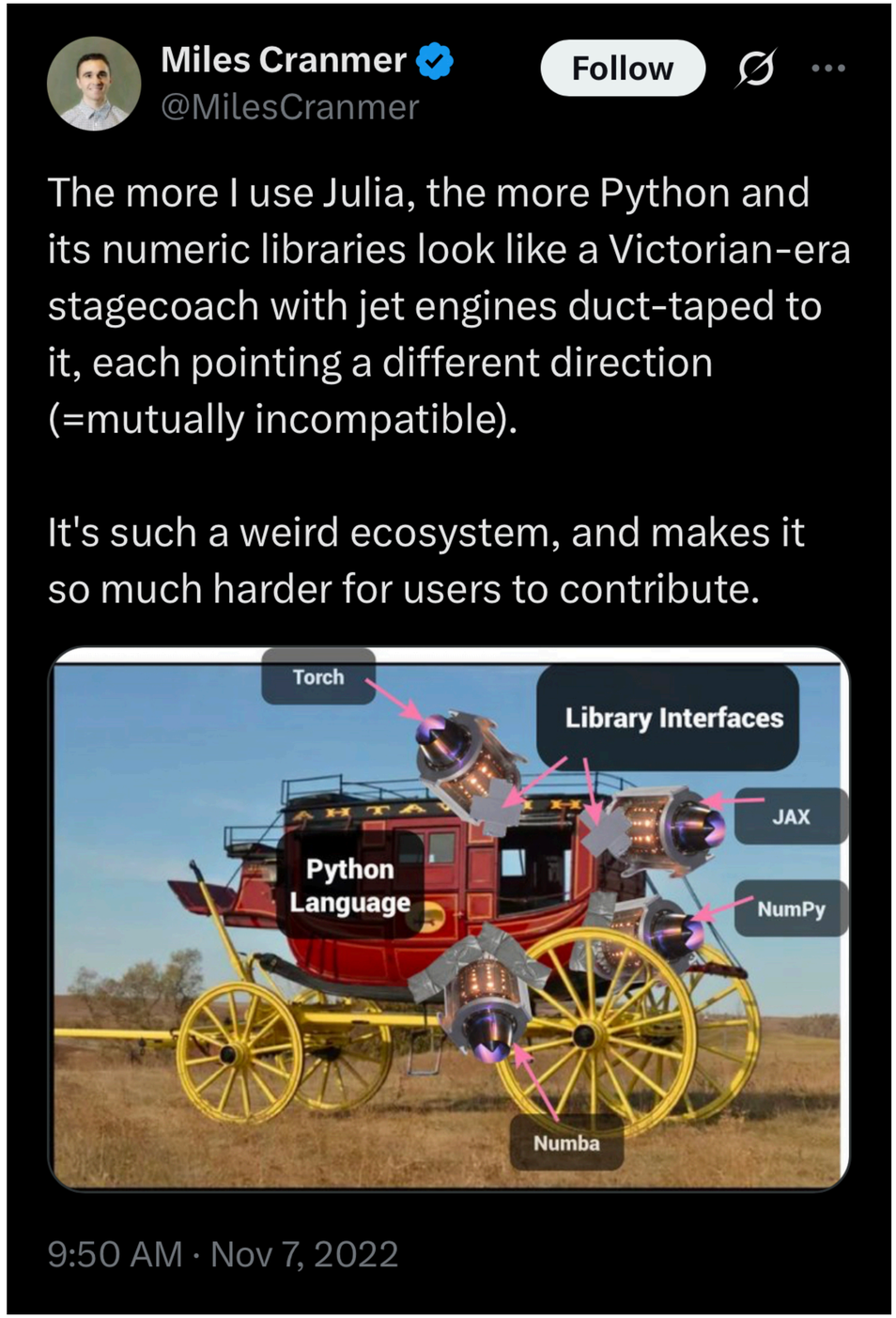
StateSpaceDynamics.jl is a comprehensive and self-contained Julia package for working with probabilistic state space models (SSMs). It implements a wide range of state-space models, taking inspiration from the SSM package written in Python by the Linderman Lab. This package is designed to be fast, flexible, and all-encompassing, leveraging Julia's speed and expressiveness to provide researchers and data scientists with a powerful toolkit for state-space modeling.

This package is geared towards applications in neuroscience, so the models incorparate a certain neuroscience flavor (e.g., many of our models are trialized as common in experiemntal paradigms). However, the models are general enough to be used in other fields such as finance, robotics, and many other domains involving sequential data analysis.

We are continuously working to expand our model offerings. If you have suggestions for additional models or features, please open an issue on our GitHub repository.

Senne et. al, submitted

w/ Ryan Senne, Zach Loschinskey



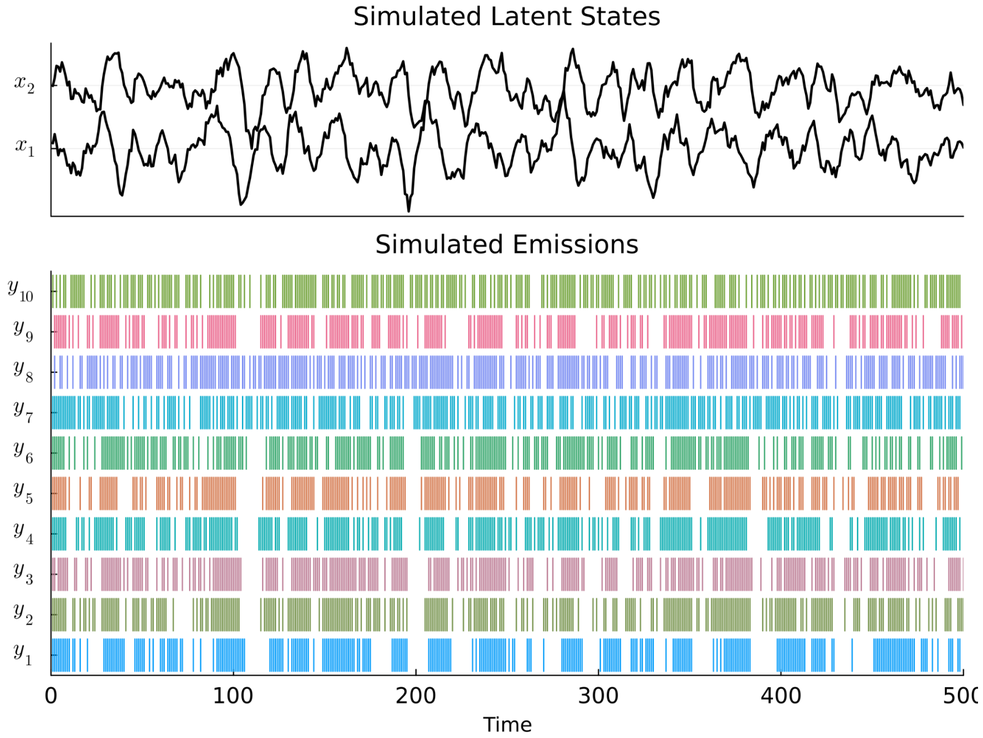
julia

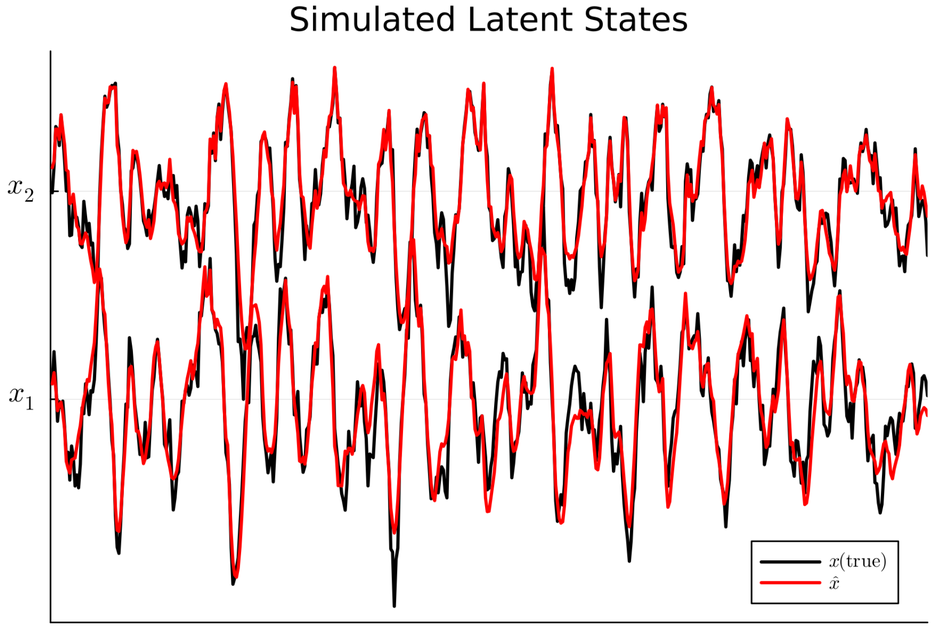
- Designed to solve the two language problem: “walks like python, runs like C”
- Developed (2009) for scientific computing, native support for complex mathematical operations. Python (1991).
- Multi-threading and multiprocessing are way easier. (JAX ✅, but now you’re stuck there).
- “Library of libraries”. Many packages developed for specific scientific domains e.g. DifferentialEquations.jl, Turing.jl, Flux.jl, which all interface easily! (e.g., neuralODE)
- World class package manager (no more unsolvable environments!)

Example: Poisson LDS

- Assume a neural population is generated through a small set of latent factors (Macke et al. 2011)

$$X_t \sim \mathcal{N}(AX_{t-1}, Q)$$
$$Y_t \sim \text{Poisson}(\exp(CX_{t-1} + d))$$

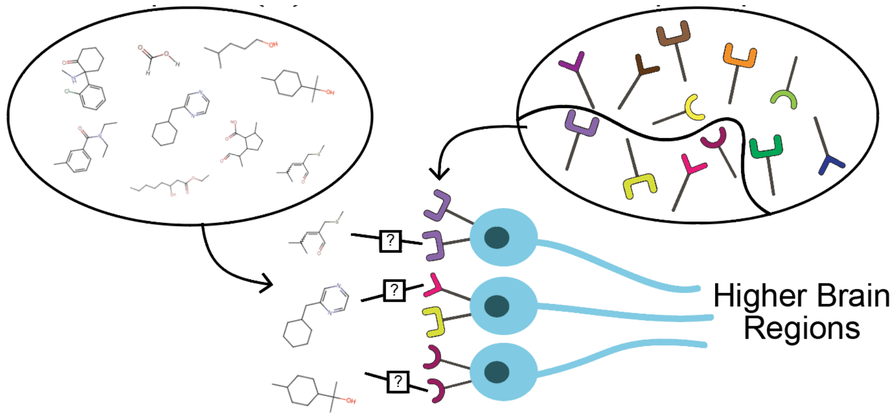




Senne et. al, submitted

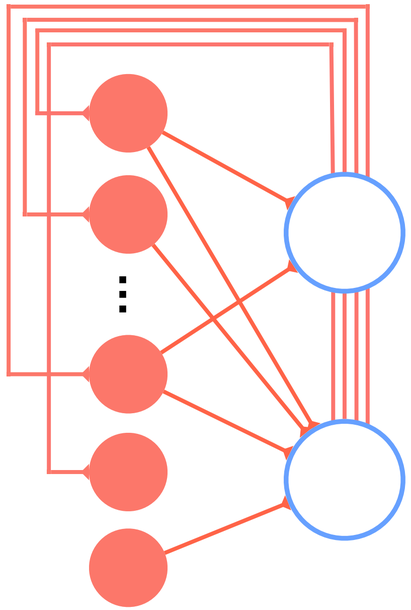
THE DEPAQ LAB @ BU

Foundation models for olfaction



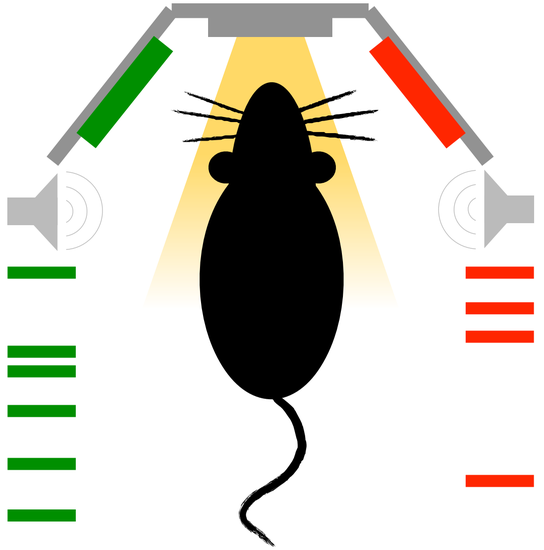
- **Best performance with protein-odor models**
- **Need better chemical context**

Spike variability from multi-task spiking networks



- **Spike variability arises due to off-manifold dynamics in multi-tasking networks**

Hierarchical models of decision-making



- **HMM-DDM captures shifting speed-accuracy tradeoff in mice**

THANK YOU!

New(-ish) lab at BU!

Grant McConachie
Darcy Zi
Ryan Senne
Brittany Ahn
Omar El Sayed
Zach Loschinskey
Tushar Arora
Halley Dante

GNN Workshop

StateSpaceDynamics.jl

Collaborations

Ben Scott (BU)
Meg Younger (BU)
Mike Economo (BU)
Mark Howe (BU)

Past Collaborators

Larry Abbott (Columbia)
Mark Churchland (Columbia)
David Sussillo (Stanford)
Jonathan Pillow (Princeton)
Carlos Brody (Princeton)
Tim Buschman (Princeton)
Rob Froemke (NYU)
Diksha Gupta (Princeton)
Michelle Insanally (Pitt)
Tim Kim (Princeton)
Thomas Luo (Princeton)
Matt Panichello (Princeton)
Lucas Pinto (Northwestern)
Kanaka Rajan (Harvard)
Abby Russo (CTRL Labs)
David Tank (Princeton)

Questions: bddepasq@bu.edu

Funding: Research Corporation for Science Advancement, Allen Frontiers Group